Uses of Theory in Randomized Field Trials

Lessons From School Voucher Research on Disaggregation, Missing Data, and the Generalization of Findings

WILLIAM G. HOWELL
PAUL E. PETERSON
Harvard University

By design, randomized field trials (RFTs) avoid many of the problems that plague observational studies, foremost among them being the introduction of selection biases. In practice, however, RFTs regularly confront other difficulties, such as chance differences between treatment and control groups and attrition from the study. To address these issues, baseline data on the variable of primary interest are essential. Theory also aids the analytic process, identifying ways in which data should be disaggregated and determining the generalizability of the findings uncovered. Theory and testing are not neatly divided enterprises. Theory informs the initial design whereas empirical findings from RFTs motivate analysts to update and occasionally abandon their theoretical priors.

Keywords: *randomized field trials; school vouchers; sample attrition*

After decades of bitter conflicts and frequent backtracking, astronomers now think they know the future of the universe. For years, they postulated that the universe eventually would collapse in on itself, ending in a fiery blaze. Then it was to remain ensconced in a steady state; then to expand continuously, although at a declining rate. Now, according to recent experiments, the universe seems to be rushing onward and outward at an ever-increasing pace, as all matter and energy dissipates into a void.

Experiments, together with newly possible telescopic observations, propagated many of these developments. Experiments revealed that not enough matter was visible to keep the universe together; so theorists, building on Einstein's constant, invented enough dark matter to slow the expansion. Experiments then revealed accelerating distances between galaxies, and so theorists invented dark energy to counteract dark matter. Even as experiments now indicate that the universe is expanding ever outward, cosmologists are imagining the possibility

Authors' Note: *This article was prepared for the Institution for Social and Policy Studies, Yale University.*

AMERICAN BEHAVIORAL SCIENTIST, Vol. 47 No. 5, January 2004 634-657
DOI: 10.1177/0002764203260153

of multiple universes, still unseen, that may help sustain our belief in eternal life, or at least provide fodder for an eternal dialogue between theory and experimentation.

Social scientists participate in a similar conversation between theory and experimentation. Theory directs the analytic process, ordering and assigning meaning to findings; the findings themselves, meanwhile, regularly require that theoretical intuitions be updated. The process is wholly dynamic, with theory motivating and guiding research, and findings from experimentation corroborating, rejecting, or forcing a modification of theory.

Philosophers of science have made this point time and again, although usually giving theory pride of place. Karl Popper (1959) argued vehemently against the ideas of inductive logic. Empirical science, he insisted, requires "putting forward and testing theories" (Boyd, Gasper, & Trout, 1999, p. 99). Facts, as such, are not intrinsically meaningful; they acquire meaning when they test aspects of theory. Alone, fact gathering does not advance scientific knowledge. But when ideas are empirically tested, when theoretical propositions are subject to the uncompromising and unapologetic judgment of data (appropriately collected and analyzed), science muddles onward.

Popper (1999, p. 99) goes on to "distinguish sharply the process of conceiving a new idea, and the methods and results of examining it logically" (Boyd, Gasper, & Trout, 1999, p. 99). Accordingly, analysts should collect data only after theories are sufficiently developed and predictions appropriately derived. Intellectual honesty presumably requires that scientists establish the logic of their theories before peeking at the results of their data. The best experiments, according to Popper, are those designed to test particular hypotheses.

The demarcation of theory and experimentation, however, can be overdrawn. Indeed, we are not convinced that the processes of theory building and experimentation can (or should) be sequestered from one another. Two objections, from our perspective, stand out. First, Popper overemphasizes the temporal succession of theory construction and testing. Just because theoretical claims are specified in advance of experimentation, as Popper recommends, does not necessarily make them more valid—or more useful—because today's ex-post-theoretical justifications are tomorrow's working assumptions awaiting falsification. Second, from a sociological perspective, Popper overlooks the symbiotic relationship shared between theory and experimentation. Theory, we suggest, informs the conduct of experiments, from the construction of initial hypotheses to the development of research designs to the diagnoses of methodological problems and, ultimately, to the generalization of results. If Popper's distinction between constructing and testing scientific hypotheses may be useful analytically, in practice, the enterprises are so interconnected and so interdependent as often to be indistinguishable from one another. Theory emerges from experimental research just as it motivates it.

The task of differentiating theory from empiricism falls as much on a discipline as any particular research team. Recall the example with which we began:

With new experimental findings and observations, individual astronomers, physicists, and astrophysicists updated (again and again) their thoughts about the universe's future. Philosophically, it may be useful to neatly separate theory from experimentation—distinguishing, for example, experimental from theoretical physics and astrophysics. In practice, though, scholars reconsider theoretical first principles in light of new empirical findings, just as they redirect experimental research to test new theoretical insights.

Randomized field trials (RFTs), the topic at hand, represent just one form of experimentation. They lie midway between classic experiments, which fix all variables save the one of interest, and natural experiments, which take advantage of exogenous changes in the real world. Similar to other types of experiments, RFTs limit selection biases by randomly assigning subjects to treatment and control conditions. Because they occur outside of a laboratory setting, however, RFTs do not confront as many concerns about external validity as do classic experiments. And because they involve the deliberate manipulation of social processes, RFTs lend analysts a degree of control unavailable in most natural experiments. Still, natural experiments and observational studies are not to be discarded, if only because they help elucidate the applicability of results from RFTs to different populations and geographic regions.

Although the subject matter may lack the panache and grandiosity of expanding and collapsing universes, scientific investigations of education policy demand much the same logic of inquiry. This article examines the roles of theory and experimentation in a randomized field trial of a small New York City school voucher program. The first section briefly describes the intervention and the procedures used to evaluate it.[1] The second section illustrates the importance of drawing on theory to determine how, and whether, data ought to be disaggregated for subpopulations. The third section underscores the value of theory when addressing missing data problems that arise in most research enterprises. The fourth section emphasizes the need for both theory and observational data when generalizing findings beyond particular settings.

SECTION 1: THE SCHOOL VOUCHER EVALUATION

School vouchers, which provide tuition subsidies for students interested in attending a private school, represent one of the most controversial policy reforms in education today. By challenging public school monopolies, shifting powers from the state administrators to parents, and reshaping school assignment procedures, school vouchers have captured the imaginations and mobilized the opposition of some of the most prominent interest groups in America: the American Federation of Teachers, the National Education Association, the American Civil Liberties Unions, and the National Association for the Advancement of Colored People. In every branch of government, at both the state and federal levels, a decade-long battle has been waged over school vouchers' rightful place in the education landscape. And if the Supreme Court's recent decision on the Cleveland voucher program is any indication,[2] political fights are likely to continue for some time to come.

Essential points of fact also remain unresolved. Indeed, until the mid-1990s, very little was known about whether school vouchers actually improved student learning. Although numerous observational studies compared the achievement levels of public and private school students (more on these below), serious methodological concerns lingered. Because private schools charge tuition and retain considerable discretion when admitting (or not admitting) students, analysts were forced to compare self-selected populations. Although it may be possible to control for observable student and family background characteristics, it is extremely difficult to parse the influence of one intangible factor: the willingness and ability of parents to pay the costs (financial and otherwise) of a private education and all that this indicates about the importance they place on their child's schooling.

The best way to overcome selection biases, of course, is to randomly assign students to public and private schools, because only then can analysts be sure that observed differences are due to the schools students attend and not the social and economic advantages they bring with them. For a variety of reasons, however, the conditions necessary to perform a high-quality randomized field trial of school vouchers never arose—at least until 1997, when a group of philanthropists established the School Choice Scholarships Foundations (SCSF) in New York City.

In the spring of 1997, SCSF invited applications from students interested in vouchers worth as much as $1,400. Students in Grades K-4 who attended a public school and who were eligible for participation in the free lunch program qualified for a voucher. More than 20,000 students expressed an interest in the voucher. Rather than hand out vouchers on a first-come, first-served basis, program administrators opted to randomly award them by means of a lottery. The lottery was held in May 1997, and that fall, recipients attended private schools.

Approximately 1,200 students were offered vouchers, which were initially guaranteed for 3 years. During the program's 1st year, 74% of families offered vouchers actually used them to send their children to private schools; after 2 and 3 years, 62% and 53% of the treatment group continued to attend private schools, respectively. Meanwhile, in all 3 years, a small percentage of the control group (less than 5%) found alternative funding sources to pay the costs of a private education.

Because subjects were randomly assigned to treatment and control conditions, the procedures used to evaluate the SCSF program conform to those in randomized field trials. The evaluation team collected baseline data prior to the lottery, administered the lottery, and then collected follow-up information 1, 2, and 3 years later. This section reviews the steps taken to collect the relevant information.

BASELINE DATA COLLECTION

During the eligibility verification sessions attended by voucher applicants, students in first grade and higher took the Iowa Test of Basic Skills (ITBS) in reading and mathematics. The sessions were held during the months of February, March, and April immediately prior to the voucher lottery and generally lasted about 2 hours. The 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. The producer of the ITBS graded the tests.[3]

While children were being tested, accompanying adults completed surveys that asked about their satisfaction with their children's schools, their involvement in their children's education, and their demographic characteristics. This article considers only test-score outcomes for students with baseline and follow-up data. Other outcomes, as reported by parents and students, are reported elsewhere (Howell & Peterson, 2002).

More than 5,000 students attended baseline sessions in New York City. Mathematica Policy Research administered the lottery in May 1997; SCSF announced the winners. Thereafter, approximately 1,000 families were selected at random from those who did not win the lottery to comprise a control group of approximately 960 families.[4]

Because vouchers were allocated by a lottery, those offered scholarships are not expected to differ significantly from members of the control group (those who did not win a scholarship). Baseline data confirm this expectation (see Peterson, Myers, Howell, & Kim, 1999). Baseline test scores—far and away the best predictor of follow-up test scores, eclipsing the relative predictive power of all other demographic indicators[5]—for treatment and control group members were 19.3 and 20.0 National Percentile Ranking (NPR) points, respectively. For these students, therefore, we can safely attribute to the programmatic intervention perceived differences between the two groups' downstream test scores.

COLLECTION OF FOLLOW-UP INFORMATION

The annual collection of follow-up information commenced in New York City in the spring of 1998. Testing and questionnaire administration procedures were similar to those that had been followed during the baseline sessions. Adult members of a family completed surveys that asked a wide range of questions about the educational experiences of their oldest child within the age range eligible for a scholarship. Students completed the ITBS and short questionnaires. Both the voucher students and students in the control group were tested in locations other than the school they were currently attending.

SCSF conditioned the renewal of scholarships on participation in the evaluation. Also, families selected to become members of the control group were compensated for their expenses and told that they could automatically reapply for a new lottery if they participated in these follow-up sessions. Overall, 82% of students in the treatment and control groups attended the Year 1 follow-up session, as did 66% in Year 2 and 67% in Year 3.

SECTION 2: ISOLATING EFFECTS

To detect programmatic effects in randomized field trials, some may argue, one can ignore theory and simply compare outcomes for treatment and control group members. Where positive differences arise, the intervention appears effective; where negative differences arise, the intervention may be counterproductive; and absent any differences at all, the intervention is probably innocuous. To evaluate a randomized field trial, no prior expectations are presumably needed and, gratifyingly, none therefore impede.

Analysts that disregard theory, however, do so at their own peril. Without some insight into the underlying data-generating process, analysts may overlook important differences among subpopulations. Unless treatment effects apply uniformly—and they rarely do—analysts may falsely conclude that an intervention is benign when in truth it significantly helps some subjects or hurts others. Within the context of medical trials, men may benefit greatly from a pill, whereas women do not; the elderly may respond differently than the young; and for people with certain kinds of preexisting conditions, a treatment may be devastating. To isolate the appropriate comparison groups, analysts must surmise how medical interventions interact with the physiology of different subjects; that is, they require some sense for how the treatment actually works. Without some theoretical grounding, analysts may fail to disaggregate findings for particular groups who respond to treatment in unique ways.

When considering the impact of school vouchers on student test scores, it is useful to begin with a basic theory that accounts for the educational choices Americans make when selecting a place of residence. Families pick schools when they decide where to live. As such, school choice is not an abstract vision of a potential future but rather a deeply embedded feature of contemporary practice. School vouchers do not so much introduce choice in education as disrupt its dependence on housing markets.

Given its prevalence, school choice by residential location should have varying effects on different subpopulations. Those willing and able to pay the price of moving to select neighborhoods reap the educational benefits of better schooling.[6] Low-income families, meanwhile, lack the earning power to buy into districts with quality schools that suit the particular needs of their children. Quite the opposite, they often can afford a home or apartment only because it is located in poorer neighborhoods with inferior schools.[7]

In several ways, African Americans suffer most from this arrangement. They have lower incomes and less wealth (Davern & Fisher, 2001, pp. 70-71; U.S.

Bureau of the Census, 2000, Table 744, p. 470), are less likely to obtain a mortgage and own a home (Bullard, 1994, p. 194; Simmons, 1997, p. xvii; U.S. Department of Commerce, 1999, Table 2-1, p. 42),[8] and are more likely to live in poorer neighborhoods with greater social problems (Bostic & Surette, 2000). Furthermore, African Americans are more likely to face discrimination in housing markets, further disabling their ability to gain access to quality public schools (Munnell, Browne, McEneaney, & Tootell, 1992; Munnell, Tootell, Browne, & McEneaney, 1996).[9] The net results of economic forces and racial discrimination are highly segregated housing markets, especially within urban regions (Bullard, Grigsby, & Lee, 1994, p. 4; James, 1994, p. 99). Such trends obviously affect African Americans' ability to exercise school choice by residential selection. Precisely because they have fewer options about where to live, they have fewer choices about where to educate their children.

New forms of choice may be expected to have, in the short run, differential impacts on subpopulations, depending on whether families benefit from existing choice arrangements. Among those who enjoy a broad array of education options, the marginal benefits of school vouchers should be quite small. But where residency patterns yield poor educational options, the impacts of voucher programs may prove relatively large. Ethnicity, as such, may critically determine who benefits from vouchers.

A large body of observational data bolsters the claim that educational gains associated with switching from public to private schools are concentrated among African Americans (Evans & Schwab, 1993; Figlio & Stone, 1999; Grogger & Neal, 2000; Neal, 1997; Rouse, 2000).[10] Jeffrey Grogger and Derek Neal (2000), for instance, argue that "urban minorities in Catholic schools fare much better than similar students in public schools" (p. 153), whereas the effects for urban Whites and suburban students generally are "at best mixed" (p. 153). Moderating a debate in a special edition of the *Sociology of Education*, Christopher Jencks (1985) determined that "the evidence that Catholic schools are especially helpful for initially disadvantaged students is quite suggestive, though not conclusive" (p. 134). None of these scholars offers a comprehensive theory for why urban minorities generally, and African Americans in particular, benefit from a private education. All, though, identified an empirical regularity quite consistent with a theory of residential choice.

Hispanics and African Americans constituted more than 90% of the population in the New York City voucher experiment. Clearly, the residency patterns of Hispanics and African Americans do not differ as much as those of African Americans and Whites. In urban centers nationwide, however, Hispanics are less likely to be denied a home loan than are African Americans; levels of residential segregation for Hispanics trail those of African Americans, and Hispanics are more likely than African Americans to move to a community because of the quality of its public schools (see Frankenberg & Lee, 2002; Howell & Peterson, 2002, pp. 23-27). In New York City, African Americans are more isolated from Whites than are Hispanics; poor African American children attend more segregated schools than do Hispanics; and average neighborhood disparities in the median household incomes of Whites and African Americans are greater than those between Whites and Hispanics.[11] To the extent that African Americans live in more segregated neighborhoods with public schools that do an inferior job of addressing their individual needs, they may benefit relatively more from an opportunity to attend a private school.

Not all students who were offered vouchers in New York City attended private schools, and not all students in the control group remained in public school. Because the decision to actually use vouchers is nonrandom, one cannot simply compare public and private school parents to estimate programmatic impacts. To recover consistent estimates of the impacts of actually using a voucher, we rely on the lottery (which randomly offered vouchers to families) as an instrument for private school attendance. We estimate the following two-stage, least squares regression:

$$\begin{aligned} P_{\mathrm{t}} &= \alpha_0 + \alpha_1 V + \alpha_2 Y_{0R} + \alpha_3 Y_{0M} + \Sigma\alpha_{\mathrm{i}} L_{\mathrm{i}} + \mu_1 \\ Y_{\mathrm{t}} &= \beta_0 + \beta_1 P_{\mathrm{t}} + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \Sigma\beta_{\mathrm{i}} L_{\mathrm{i}} + \mu_2 \qquad (1), \end{aligned}$$

where Y_t is each student's total achievement score on the Iowa Test of Basic Skills expressed in NPR points,[12] and the subscript t denotes the year the student completed the follow-up test (either 1, 2, or 3). The total achievement score is a simple average of the math and reading components.[13] V is an indicator variable for whether an individual was offered a voucher, P_t is an indicator variable for attendance at a private school for t years, and L_{i} are lottery indicators.[14] Y_{0R} and Y_{0M} are the baseline reading and math scores.[15] The β_1 coefficient represents the estimated impact of switching from a public to a private school on student test scores. Specifically, β_1 represents the difference in test scores between those students who used vouchers to attend a private school and those who would have used a voucher had they been offered one.

The findings in Table 1 suggest that impacts do, in fact, vary by ethnicity.[16] Overall, students who used vouchers did not score any higher, or any lower, than their peers who remain in public schools. But for African Americans, substantial differences are observed.[17] African Americans in private schools who were retested after 1, 2, and 3 years scored, on average, 3.4, 3.2, and then fully 7.8 NPR points higher than their peers in public schools on the combined reading and math portions of the Iowa Test of Basic Skills.[18] Meanwhile, no differences are detected for members of other ethnic groups, most of whom are Hispanic. As a residential theory of choice would indicate, substantial differences in outcomes are logged for students from different ethnic backgrounds.

There are, of course, any number of ways to disaggregate test score data—by student grade level, ability, immigrant status, age, or mother's education. Indeed, we ourselves have examined effects for all of these subpopulations.[19]

TABLE 1: Simple Estimates of Impacts of Switching to a Private School on Test Score Performances

Test Score Performance	*Year 1 (%)*		N	*Year 2 (%)*		N	*Year 3 (%)*		N
All students	1.14	(1.09)	1,449[a]	0.18	(1.28)	1,199	1.37	(1.52)	1,250
African Americans	3.35**	(1.48)	622	3.20*	(1.72)	497	7.79**	(2.23)	519
All other ethnic groups	–0.31	(1.62)	812	–0.82	(1.85)	699	–1.64	(2.14)	729

NOTE: Bootstrapped standard errors that are robust to intrafamily correlations are reported in parentheses. Weighted two-stage least squares regressions were performed; treatment status was used as the instrument. All models control for baseline math and reading test scores and lottery indicators. Impacts expressed in terms of national percentile rankings for composite (math and reading combined) test scores. In Howell and Peterson (2002), we report regular ordinary least squares (OLS) standard errors. The argument for bootstrapping rests on 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. For sake of consistency in this volume, we report bootstrap standard errors.
a. The number of African Americans and members of other ethnic groups do not sum to the total number of students because of missing values on the ethnicity variable.
*p ≤ .10. **p ≤ .05., two-tailed.

Rather than turning to theory for guidance, one might instead rely on blind empiricism to detect programmatic effects, cutting the analyses ever which way and letting results speak for themselves.

When selecting among comparisons, however, the analyst ultimately needs theory, because rampant empiricism cannot distinguish idiosyncratic from genuine findings. Suppose we found large and positive effects for second-and fourth-grade students but no effects for third- and fifth-grade students.[20] Should we infer that vouchers benefit members of only even-numbered grades? Obviously not. Had we discovered that second and third graders consistently benefited, but not fourth and fifth graders, we might have been more inclined to assign meaning to the results, because there is reason to expect older students to have a harder time adjusting to their new schools. Again, though, such reasons trace back to intuitions into how vouchers work, which in turn require theoretically grounded insight into the underlying data-generating process.

SECTION 3: MISSING DATA

Most randomized field trials, especially those that track low-income subjects over time, encounter missing data problems.[21] Evaluators lose track of some individuals and others refuse to continue cooperating with the study. With follow-up data only available for a subset of the entire sample, evaluators must continually diagnose the sources of attrition and assess the impacts they have the study’s internal and external validity.

TABLE 2: Response Rates for Students Taking Tests (%)

	Treatment	*Control*
African Americans		
Year 1	79.8	74.3
Year 2	67.0	55.4
Year 3	66.3	62.2
All other ethnic groups		
Year 1	82.5	77.9
Year 2	73.2	64.0
Year 3	75.0	68.1

NOTE: The numbers represent the percentage of African Americans and non–African Americans in the treatment and control groups that were tested after 1, 2, or 3 years, given that they were tested at baseline. Response rates for test scores and parental surveys differ somewhat (see Howell & Peterson, 2002).

To obtain high response rates in the New York evaluation, program operators either required or strongly urged voucher recipients to participate in testing sessions if they wished to remain in the program for the following school year. In addition, to encourage members of the control group and members of the treatment group who remained in public schools to return for follow-up testing, the evaluation team offered financial incentives and new opportunities to win a voucher. Still, substantial numbers of students were not tested at the end of 1, 2, and 3 years.

Table 2 reports the percentage of African Americans and members of other ethnicities in the treatment and control groups who completed follow-up tests after 1, 2, and 3 years.[22] For both populations, response rates are highest in Year 1 and roughly comparable in Years 2 and 3. Attrition tended to be slightly higher among African Americans than members of other ethnic groups, especially in Years 2 and 3. Finally, for all ethnic groups, attrition tended to be slightly higher in the control group than the treatment group.

If those students who were retested after 1, 2, and 3 years differed substantially from the larger population tested at baseline, then the initial randomization may have been lost, leaving essentially observational data. According to Thomas Cook and Donald Campbell (1979), the occurrence is hardly rare. "Many randomized experiments in practice move toward quasi-experiments in which pretreatment differences are to some extent confounded with treatments" (p. 360).

To the extent that attrition is nonrandom and correlated with the outcome of primary interest (test scores), then, comparisons of raw test score outcomes may be biased. To address the problem, we again turn to theory, identifying the factors that may encourage some families to attend follow-up testing sessions more than others. Depending on the character and strength of these factors, the simple estimates presented in Table 1 may underestimate or overestimate the true impacts of switching from a public to a private school.

Family background characteristics. To see whether attrition altered the composition of treatment and control groups in ways that could bias results, we begin with the generally accepted theory that wealthier, better educated, and more stable families are likely to be disproportionately represented at follow-up testing sessions. Single-parent families, especially those who work on weekends, may have had an especially difficult time attending the sessions, most of which were conducted on Saturday mornings and afternoons. Furthermore, planning and securing transportation for these sessions requires a measure of motivation and foresight, again encouraging the attendance of more advantaged members of the study.

Because those invited to participate in the follow-up studies had provided information about their family characteristics at baseline, it was possible to test the claim that attendees of follow-up testing sessions were especially advantaged. As Table 3 shows, African American respondents in the treatment group were less likely to receive welfare and tended to live in their residences for longer periods of time than did nonrespondents. Differences of comparable magnitude apply to African American respondents and nonrespondents in the control group.[23] Although no differences are observed with regard to student test scores, church attendance, religious identification, mother's education, and family size, the differences that are observed tend to suggest that selection effects favored more advantaged members of both treatment and control groups. Observation, it appears, is consistent with theoretical expectations.

As investigators in other evaluations conventionally do, we adjusted for nonresponse bias by generating yearly weights for those parents and students in the treatment and control groups who continued to participate in the study. These weights are based on logistic regressions that posited attendance at follow-up sessions as a function of demographic characteristics assembled from baseline surveys. To allow for as much flexibility as possible, separate models were estimated for treatment and control group members. These models generated a set of predicted values that represent the probability that individuals, given their baseline characteristics, would attend a follow-up session. The weights are simply the inverse of these predicted values; that is,

$$W_j = \frac{1}{F(X\beta)},$$

where $F(X\beta)$ is the model's logistic distribution function. The weights then were rescaled so that they summed to the total number of actual observations.

With these weights, we reestimated the test-score impacts, which are shown in Table 4. Given the modest differences between respondents and nonrespondents, the weights exert little influence on estimated impacts. Although the magnitude of estimated impacts are generally larger than those reported in Table 1, essentially the same patterns hold: in all 3 years, significant and positive impacts are observed for African American students who switch from public to private

TABLE 3: Characteristics of Respondents and Nonrespondents in Treatment and Control Groups

	Treatment		*Control*	
Characteristics at Baseline	*Tested in Year 3 (1)*	*Not Tested in Year 3 (2)*	*Tested in Year 3 (3)*	*Not Tested in Year 3 (4)*
African Americans				
% welfare recipients	56.4	62.0	63.0	69.3
% Catholic	16.0	24.2	15.5	14.1
% Protestant	65.2	66.6	66.3	72.1
Average composite test scores	20.6	19.2	22.4	22.2
Average family size	2.6	2.5	2.7	2.7
Average residential stability	3.8	3.6	3.6	3.7
Average church attendance	3.4	3.3	3.1	3.4
Average mother's education	2.5	2.5	2.5	2.6
All other ethnic groups				
% welfare recipients	51.4	64.8	51.3	60.4
% Catholic	78.9	72.5	77.9	72.5
% Protestant	13.3	13.8	13.1	18.8
Average overall test scores	19.6	19.7	23.3	22.8
Average family size	2.7	2.3	2.5	2.5
Average residential stability	3.7	3.7	3.8	3.6
Average church attendance	3.9	3.2	3.6	3.7
Average mother's education	2.4	2.5	2.4	2.2

NOTE: Averages refer to the mean scores of responses on the baseline parent surveys. Columns 1 and 3 refer to families that attended Year 3 follow-up testing sessions; columns 2 and 4 refer to families that did not attend Year 3 follow-up testing sessions. Only students who completed tests at baseline are included. The treatment group consists of all students who were offered a voucher and participated in the baseline study; the control group consists of all students who were not offered a voucher. Significant tests are not possible due to multiple lotteries.

TABLE 4: Estimating Test Score Impacts of Switching to a Private School Using Weighted Data

Test Score Performance	*Year 1 (%)*		N	*Year 2 (%)*		N	*Year 3 (%)*		N
All students	1.76	(1.49)	1,449[a]	0.85	(1.56)	1,199	1.52	(1.90)	1,250
African Americans	6.13**	(1.74)	622	4.16*	(2.22)	497	8.43**	(2.86)	519
All other ethnic groups	–1.97	(2.27)	812	–0.88	(2.12)	699	–3.20	(2.65)	729

NOTE: Bootstrapped standard errors that are robust to intrafamily correlations are reported in parentheses. Weighted two-stage least squares regressions were performed; treatment status was used as an instrument. All models control for baseline math and reading test scores and lottery indicators. Impacts are expressed in terms of national percentile rankings for composite (math and reading combined) test scores.

a. The number of African Americans and members of other ethnic groups do not sum to the total number of students because of missing values on the ethnicity variable.

$*p \leq .10$. $**p \leq .05$., two-tailed.

schools and no differences are observed for everyone else. Weighted estimates for African Americans are 6.1, 4.2, and 8.4 NPR points for Years 1, 2, and 3, respectively. We regard these figures as the best available estimates of the impact of switching from a public to a private school (Howell & Peterson, 2002, p. 162).

School experiences. If using background characteristics to weight cases differentially is conventional practice, methodological solutions to other aspects of the missing data problem are not. What if the factors that affect participation rates in follow-up sessions have less to do with a family's station in life and more to do with its experiences in public and private schools during the course of the evaluation? It is possible that change in academic performance over time, rather than observable baseline characteristics, affect the likelihood that different subgroups within the treatment and control groups attend subsequent testing sessions.

Consider the selection-on-treatment theory suggested to us by University of Chicago economist Derek Neal. Treatment group families who do not benefit from vouchers will tend to drop out of the study, but control group families whose children are doing ever worse in public schools return faithfully in the hopes of winning a scholarship for the coming year. If true, then observed voucher impacts are inflated.

For Neal's selection-on-treatment theory to hold, two conditions must apply: first, gains in test scores from baseline to Year 1 (2) must decrease the probability that members of the control group would attend the Year 2 (3) testing session, and second, gains must increase the probability that members of the treatment group would attend the Year 2 (3) testing session. If the differences in observed impacts on response rates for the treatment and control groups are statistically significant, estimates of programmatic effects may be biased.

The theory suggests that missing data can arise due to events that occur during the course of an evaluation. Using the baseline demographic characteristics of respondents and nonrespondents to weight the data does not address this problem, because it is the subsequent experiences of students in public and private schools that may affect their continued participation in the study. Fortunately, because data were collected over 3 follow-up sessions, it is possible to diagnose the extent of selection on treatment.

The following logistic regressions test Neal's hypothesis:

$$\Pr(A_2 = 1) = \alpha_0 + \alpha_1(Y_{1C} - Y_{0C}) + \alpha_2[(Y_{1C} - Y_{0C}) * V] + \alpha_3 Y_{0R} + \alpha_4 Y_{0M} + \mu_1$$

$$\Pr(A_3 = 1) = \beta_0 + \beta_1(Y_{2C} - Y_{0C}) + \beta_2[(Y_{2C} - Y_{0C}) * V] + \beta_3 Y_{0R} + \beta_4 Y_{0M} + \mu_2 \qquad (2),$$

where A_2 and A_3 identifies whether a student attended the follow-up sessions in Years 2 and 3, respectively; Y_{1C} and Y_{2C} are the total achievement scores at

TABLE 5: Effect of Change in Test Scores from Baseline to Year 1 and Year 2 on Students Attendance at Follow-Up Testing Sessions

	Year 2 Attendance		*Year 3 Attendance*	
	African Americans	*Other Ethnic Groups*	*African Americans*	*Other Ethnic Groups*
$Y_{1C} - Y_{0C}$	0.005 (0.013)	0.004 (0.009)		
$Y_{2C} - Y_{0C}$			0.023 (0.018)	0.000 (0.013)
$(Y_{1C} - Y_{0C}) * V$	0.004 (0.015)	0.005 (0.011)		
$(Y_{2C} - Y_{0C}) * V$			0.031 (0.021)	0.001 (0.016)
$Y_{0\text{M}}$	0.003 (0.007)	0.011* (0.006)	0.007 (0.008)	0.001 (0.001)
$Y_{0\text{R}}$	0.007 (0.005)	0.002 (0.005)	0.006 (0.007)	0.011 (0.006)
N^{a}	623	817	497	699
Pseudo R^2	0.00	0.01	0.02	0.01

NOTE: Bootstrapped standard errors that are robust to intrafamily correlations are reported in parentheses. Logit regression models were performed on unweighted data. Y_{0M} and Y_{0R} are baseline math and reading test scores. $Y_{1C} - Y_{0C}$ refers to the change in the total math and reading test scores from baseline to Year 1; $Y_{2C} - Y_{0C}$ refers to change from baseline to Year 2. $(Y_{1C} - Y_{0C}) * V$ is an interaction term between one variable that is the difference between Year 1 and baseline test scores and another variable that indicates whether a student was offered a voucher. The dependent variable is coded 1 if the student attended either the 2nd- or 3rd-year follow-up session.

a. The sample sizes here are slightly higher than those in Tables 1 and 3 because of missing values on the private school indicator.

*$p \leq .10$, two-tailed.

Years 1 and 2; and all other variables are defined as in (1). Separate models were run for African Americans and members of other ethnic groups.

Table 5 shows the results. On the whole, the signs of the coefficients point in the expected direction. Gains in test scores from baseline to Years 1 and 2 increased the probability that members of the treatment group attended the subsequent testing session and decreased the probability that members of the control group attended the subsequent session. Across the board, however, effects are not statistically significant—not for African Americans or members of other ethnic groups in either Year 2 or 3. We find no systematic support for the contention that attrition patterns among members of the treatment and control group were a function of changes in test scores.[24]

How deep is the pool? Neither weights nor imputations necessarily solve the problem of missing data. If unobservables—for example, eagerness to obtain a scholarship or automobile ownership—affect the likelihood that treatment and control group members attend follow-up testing sessions, then weights generated from baseline survey data may not fully account for nonresponse bias. To address the possibility of selection on treatment, meanwhile, we can only impute test scores for a fraction of those students who dropped out of the study. Imputing a Year 2 (3) test score hinges on the student having attended the Year 1 (2) follow-up session; without at least one follow-up test score, we have no basis on which to impute values further downstream. Prior methodological corrections, therefore, may not eliminate attrition bias.

Another way to assess attrition bias is to estimate impacts for different response rates. Not all participants attended the first testing session to which they were invited; indeed, follow-up testing sessions were conducted over several months. Those families who attended later sessions probably better resemble those who did not show up at all than do families who attended earlier sessions. Conventional theories of response would assume that stragglers look more like nonrespondents than do early birds; after all, stragglers would have been nonrespondents had evaluators not made additional efforts to test them. If their differentiating characteristics are related to student achievement, then test score impacts should vary markedly for lower response rates.

Given that we know the dates when students came in for testing, we can generate exact estimates of the impacts of attending a private school for smaller response rates. For instance, 82% of those students who were tested at baseline attended the Year 1 follow-up session. By successively dropping the portion of students who attended later testing sessions, we can readily calculate impacts for lower response rates.

If attrition is a function of students' experiences in their public and private schools, then we should expect the estimated impacts of attending a private school to increase as response rates decline. Presumably, those students who benefit most from treatment should come earlier to the testing sessions, along with those students in the control group who were performing most poorly in public schools. Impacts of attending a private school, then, should be larger for lower response rates. The differences between the two groups, however, should attenuate (and, perhaps, switch signs) as response rates increase.

Table 6 reports the estimated impact of attending a private school for African American students for variable response rates. In each row, the last column represents the estimated impact for the full sample of African American students who attended testing sessions. Prior columns provide estimates of impacts for lower response rates, based on when students came in for testing.

In all 3 years, rather than declining as response rates increased, the positive estimated impacts grew in magnitude. Had we stopped testing African American students in Year 2 after the first 30% of the sample showed up, we would have recovered almost exactly the same findings that we observed for the full sample—the point estimate for the first 30% of students to be tested was 3.8 NPR points, and it was 4.2 for the full sample. Differences in Years 1 and 3 are more dramatic. Moving from a 30% response rate to the full sample, the estimated test score impact of attending a private school increased by 3 to 4 NPR points.[25]

The findings for non–African Americans appear slightly more stable. In Years 1 and 2, estimated impacts for the first 30% of non–African Americans to attend follow-up sessions were roughly 2 NPR points higher than the estimated impacts for the full samples. Of interest, in years 2 and 3, a significant and

TABLE 6: Estimated Impacts of Attending a Private School for Latinos and African Americans for Variable Response Rates

	Percentage of Respondents Attending Follow-Up Sessions					
	30	*40*	*50*	*60*	*70*	*Full Sample*
African Americans						
Year 1 impact	3.27	2.65	4.28**	5.30**	5.65**	6.13**
	2.08	(1.79)	(1.74)	(1.72)	(1.65)	(1.74)
Year 2 impact	3.83	3.97	4.57**	4.45**	[a]	4.16**
	(2.88)	(2.62)	(2.33)	(2.23)		(2.22)
Year 3 impact	4.50	8.08**	6.28**	8.08**	[a]	8.43**
	(3.43)	(3.15)	(2.94)	(2.89)		(2.86)
All other ethnic groups						
Year 1 impact	–0.86	–2.56	–1.98	–2.87	–2.54	–1.97
	(2.68)	(2.80)	(2.84)	(2.41)	(2.18)	(2.27)
Year 2 impact	–5.61**	–2.73	–1.61	–1.85	[a]	–0.88
	(2.66)	(2.21)	(2.17)	(2.12)		(2.12)
Year 3 impact	–1.58	–4.62	–3.25	–4.50*	[a]	–3.20
	(3.41)	(3.52)	(2.84)	(2.68)		(2.65)

NOTE: Bootstrapped standard errors that are robust to intrafamily correlations are reported in parentheses. Weighted two-stage least squares regressions were performed; treatment status was used as an instrument. Differential response rates were calculated by including in the analysis only the relevant percentage of students to initially attend testing sessions. Impacts are expressed in terms of national percentile rankings for composite (math and reading combined) test scores. All models control for baseline test scores and lottery indicators.
a. Full samples in Years 2 and 3 had less than 70% response rates.
***p* ≤ .05., two-tailed.

negative impact turns up for lower response rates. Had we tested 30% of non–African Americans in Year 2, or 60% in Year 3, we would have concluded that attending a private school negatively affected student test scores.

As response rates increase, our assessments of the efficacy of school vouchers generally improve. In all 3 years, the observed positive impacts for African Americans increase in magnitude as rising proportions of students are brought in for follow-up testing. It is impossible to know whether even larger positive gains would have arisen for African Americans had we managed to retest even more students. These estimates do suggest, however, that our findings are probably not an artifact of selection-on-treatment effects. Members of the treatment group who benefited most from attending a private school and members of the control group who suffered most from remaining in a public school were not among the first to attend follow-up testing sessions. To the contrary, if observed patterns hold for higher response rates, then we actually have underestimated the true gains associated with switching from a public to a private school.

Where do these investigations of attrition bias leave us? Depending on which theory we call on, we can push the estimated test score impacts by one or two

NPR points in either direction. These differences are sufficiently small to conclude that attrition bias in the New York City voucher program probably did not lead to a gross over- or underestimation of the true test-score impacts of switching from a public to a private school.

SECTION 4: GENERALIZING FINDINGS

In a district with more than a million students, SCSF offered vouchers to just 1,000. These vouchers tended to be quite small, never exceeding $1,400. The evaluation halted after just 3 years. By any standard, the New York City voucher program constituted a small-scale educational intervention.

To what larger populations might the estimated impacts apply? Had vouchers been worth thousands of dollars rather than hundreds, would students have gained access to a wider array of private schools? And had they, would they have benefited academically? What are the longer-term impacts of switching from public to private schools? To such questions, one returns to theory and observational studies for guidance.

If a differentiated theory of residential choice is correct, then we should expect to see comparable gains for African Americans in larger voucher programs conducted in other metropolitan regions. Indeed, gains should persist until an expanded array of public schooling options are awarded to African American students. As such, we and our colleagues have recommended the initiation of a larger scale voucher experiment in a major metropolitan region with large concentrations of African Americans (Howell & Peterson, 2002, pp. 207-208).

Of course, when bringing a social intervention to scale, much can change. Public schools may face new incentives to improve, effectively negating the marginal benefits of attending a private school. Private schools, likewise, may find it in their interest to better accommodate the particular needs and interests of Latino students who, in smaller programs, did not demonstrate any achievement gains from using vouchers. In larger voucher programs, parents may find it easier to select a school that adequately matches the individual needs of their child. Furthermore, new private educational entities may emerge to serve particular populations of students.

Other outcomes of larger scale voucher programs may be deleterious. In the smaller New York City program, we observed only moderate levels of skimming. But with greater application pools, private schools that are not required to admit students at random may select out the best and brightest and send the less fortunate back to their neighborhood public schools, effectively reinforcing the social inequalities that pervade American education. With regard to test scores, outcomes may vary according to voucher programs' size and location. Private school systems, at least initially, may be ill equipped to deal with a major influx of new students. And voucher programs in other cities, with fewer private schooling options and less established Catholic dioceses, may not replicate the gains observed in New York City.

Much may depend on selection mechanisms into voucher program. After learning that they did not receive a voucher, New York City families in the control group may have substituted educational inputs (e.g., tutoring) that are less expensive than a private school tuition—in which case, our findings underestimate the true impact of switching from a public to a private school. On the other hand, the New York City program may have attracted only those few families whose children would benefit greatly from a private education—in which case, our findings do not transfer to larger voucher initiatives.

We do not pretend to know how a large scale voucher initiative will impact the education of children in public and private schools. Although cautious optimism comes from national observational studies that demonstrate educational gains for African Americans attending private schools and from RFTs conducted in New York City and elsewhere (Evans & Schwab, 1993; Figlio & Stone, 1999; Grogger & Neal, 2000; Howell & Peterson, 2002; Neal, 1997), the interplay of theory and empirical testing never appropriately ceases. Experimental research must continually be directed to those areas of inquiry where theory presides. Once discovered, new findings will help refine or reject competing claims about the efficacy of different social interventions.

SECTION 5: SOME CONCLUDING REMARKS ON THEORY AND EXPERIMENTATION

A fiery finale to the universe made sense to scientists schooled in the book of Revelations. For centuries, dark matter and dark energy were too preposterous a set of concepts to have had serious theoretical traction—until experimentation and observation demonstrated that the observable mass could neither hold the universe together nor induce increasing acceleration. If facts have little meaning apart from theory, then it is equally true that theory cannot progress without experimentation and observation.

This simple truth resonates. When we began our research, we had a simple minded view of vouchers, one shaped by a theory of markets. When subjected to competitive pressures, this theory suggests, schools should respond by offering better products at lower costs. Because parents can choose schools that best address the needs of their child, bad schools, presumably, will lose customers, unless they quickly find ways to adapt and improve. Good schools, meanwhile, will flourish, and over time, other schools will replicate their practices. As such, vouchers should introduce flexibility and autonomy into public education and increase productivity by forcing schools to compete for customers.

Assuming that the voucher intervention was too small to induce systematic changes within the public and private sectors—changes, incidentally, that conceivably could negate observed achievement differences between treatment and control groups—we expected students who gained access to a competitive private education market to benefit academically. Nothing in market theory, however, suggested that the magnitude of the impact should systematically vary by ethnicity. But when analyzing whole populations, we observed no differences between those who used vouchers to attend area private schools and those who remained in public school.[26]

Quite by accident, while performing simple diagnostic tests in Year 2, we discovered that African Americans consistently were posting achievement gains in New York and other cities where randomized field trials were conducted.[27] The sheer robustness of the empirical findings forced us to rethink our initial assumptions about how vouchers would work, and only then did we develop a more nuanced theory of educational choice based on residential selection. The theory recognizes the power of market forces but places special emphasis on the varying capacities of families to exercise choice within the existing system—doing so, it provides a finer account of what happens to poor students in urban environments who switch from public to private schools.

Future randomized field trials assuredly will generate new findings that require accommodation, and only by combining these results with those from large-scale observational studies will we know whether a theory of choice based on residential selection can prove its scientific mettle. The process of testing and fine-tuning theory is never complete. As Donald Campbell and Julian Stanley (1963) noted some 40 years ago, what we know at any given moment represents only the "cumulation of selectively retained tentatives" (p. 4). Due to the strength of their designs, however, randomized field trials and other experimental research furnish findings that assuredly accelerate the pace at which social scientific theory advances, shifting today's tentatives out of the realm of fanciful guesswork and into that of informed understanding.

NOTES

[1] The many groups and individuals who assisted with the evaluation are acknowledged in Howell and Peterson (2002). Here, we wish to thank as well those who have provided comments on the articles included in this volume, including Alan Altshuler, Christopher Berry, David E. Campbell, Morris Fiorina, Alan Gerber, Donald Green, Jay Greene, Erik Hanushek, Frederick Hess, Caroline Minter Hoxby, Martin West, and Patrick Wolf. Howell and Peterson (2002) also includes findings (test scores and otherwise) from voucher experiments in other cities. Also, see Howell, Wolf, Campbell, and Peterson (2002) and Peterson, Howell, Wolf, and Campbell (2003).

[2] *Zelman v. Simmons Harris* (536 U.S. 639 [2002]).

[3] The assessment used in this study is Form M of the Iowa Tests of Basic Skills, Copyright 1996, by The University of Iowa, published by The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, IL 60143-2079. All rights reserved.

4. Exact procedures for the formation of the control group are described in Hill, Rubin, and Thomas (2002).

5. 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 3 outcome years. Their correlations, however, never exceed 0.25.

6. Economists have shown that the quality of a community's public schools informs the value of its property. See, for example, Black (1999); Hayes and Taylor (1996); and Bradbury, Mayer, and Case (2001).

7. For one study on the link between school quality and housing prices, see Weimer and Wolkoff (2001).

8. African Americans also have, on average, just $29,000 of equity in their homes, as compared to $36,000 for Hispanics and $50,000 for Whites (see Simmons, 1997, Table 1.37, p. 58).

9. Subsequent studies have criticized Munnell, Browne, McEneaney, and Tootell (1992) and Munnell, Tootell, Browne, and McEneaney (1996) (see, e.g., Bostic, 1997).

10. Other studies finding positive educational benefits from attending private schools include Coleman, Hoffer, and Kilgore (1982) and Chubb and Moe (1990). Critiques of these studies have been prepared by Goldberger and Cain (1982). One experimental study also found positive impacts. In Milwaukee, positive impacts of vouchers on student test scores were observed, most clearly after 3 and 4 years (Greene, Peterson, & Du, 1998), but in this randomized field trial, baseline test scores were available for only 29% of the voucher students and 49% of the control group—just 83 students after 3 years and 31 students after 4 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 3rd year, 112 in the 4th year), regardless of whether baseline information was available (pp. 345-348). All results were positive, although 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)

11. One indicator points in the opposite direction of these trends. Poor Hispanics in New York City are 5 percentage points less likely to own a home than are poor African Americans. These findings are available at http://mumford1.dyndns.org/cen2000/data.html.

12. For ease of interpretation, we report impacts in terms of National Percentile Ranking (NPR) points. The results do not change substantively when using National Curve Equivalents or raw test scores.

13. Because it is based on a larger number of test items, the total achievement score is likely to generate more stable estimates than are reading and math scores estimated separately (see Krueger, 1999). Indeed, standard errors for composite test scores are 15% to 20% lower than those for reading or math separately. For similar treatment of test scores, see Krueger (1999).

14. For the 1st year's analysis, P denotes whether an individual attended a private school for the entirety of the school year. For the 2nd year's assessment, P denotes whether an individual attended a private school for both years. In all three cities, less than 3% of the students in the treatment sample attended a private school for only 1 of the 2 years.

15. For the sample of students with baseline test scores, the inclusion of controls for baseline test scores in the model does not materially affect the magnitude of the estimated impacts of attending a private school; it does, however, substantially improve the efficiency of the estimates.

16. Estimates here differ slightly from those originally reported because Mathematica Policy Research (MPR), the firm responsible for data collection, after certifying an original set of weights and lottery indicators in 2002, revised them in 2003.

17. African American students are identified by the mother's ethnicity. See Peterson and Howell (2003, 2004) for estimated impacts for different definitions of African American.

18. The estimates in Table 1 do not adjust for attrition bias or particular aspects of the baseline lotteries. Baseline weights were calculated to adjust for the fact that students from underperforming public schools had a higher chance of being offered a voucher.

19. Only with regard to ethnicity did we find consistent effects over time and across cities.

20. In fact, we found positive effects for students with baseline test scores in all grades in Year 3.

21. Twenty-four African American students (or 10.6% of the sample) in grade 1, 34 (12.9%) in Grade 2, 21 (8.9%) in Grade 3, and 25 (13.6%) in Grade 4 had missing baseline test scores. All 245 African American kindergartners had missing baseline test scores. According to the original research proposal, MPR 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., 1996)

After the lottery was held, MPR reported that administrative procedures were not fully executed according to plan because some students for whom no baseline test scores were available nonetheless were given a chance to win a voucher. Also, MPR did not make available test score data to the propensity score matching team until after their work was completed, causing problems with the construction of the control group (Barnard et al., 2003, p. 301).

22. Weighted, 2SLS regressions estimated where treatment status is used as an instrument. Estimates of private school impacts compare those students who attended a private school for 3 years to those students who did not. If students benefited from attending a private school for 1 or 2 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 affects student achievement, then this model will underestimate the true impact of consistent private-school attendance.

It is preferable to estimate the impact of actually attending a private school rather than the impact of being offered a voucher. The latter impact, known as intent-to-treat (ITT), is estimated by ordinary least squares (OLS), the former impact by a two-stage model (2SLS), which uses the randomized assignment to treatment and control conditions as an instrument for private school attendance.

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 of 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 noncompliance among treatment and control group members were substantial in New York, OLS estimates are considerably lower than the 2SLS estimates we report above.

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 1, 2, or 3 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. As Barnard et al. (2003) point out, "one could argue that the effect of attending [a private school] . . . is the more generalizable, whereas the [effect of offering will change] . . . if the next time the program is offered the compliance rates change (which seems likely!)" (p. 321).

In addition, the OLS model does not provide 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. Because the New York voucher program was small, privately funded, initially limited to 3 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.

2SLS estimates could be biased when the voucher utilization is erroneously measured. The direction of the bias, however, remains unclear. There is no reason to expect measurement error for the treatment group because administrative records were used to identify students who were using the voucher to attend a private school. And for the control group, all students were assigned to public schools unless information reported by the parent indicated otherwise. Because some of the students in the control group for whom attendance data were missing may well have been enrolled in private schools, and because 2SLS estimates increase, relative to OLS estimates, in direct proportion to the percentage of control group members who attend private schools, recovered estimates of attending a private school appear to be downwardly biased. However, this remains uncertain inasmuch as measurement error arising from nonresponse is correlated with the instrument employed may introduce additional bias. For similar use of 2SLS estimations, see Gerber and Greene (2000) and Krueger (1999).

23. Comparable results hold for Years 1 and 3.

24. In Howell and Peterson (2202), where OLS standard errors were estimated instead of bootstraps, significant effects are observed for African Americans in Year 3. To further investigate the extent of selection-on-treatment bias, we imputed Year 3 test scores for the 78 African American students who attended the baseline and Year 2 testing sessions but failed to show up in Year 3. Imputations were based on students' treatment status, baseline test scores, test score changes between baseline and Year 2, and the Year 3 weights. Having done so, observed effects remained positive and statistically significant.

25. Comparable findings arise when estimating impacts using unweighted data.

26. Of course, by achieving comparable test scores at significantly reduced costs, private schools demonstrate efficiency gains, just as market theory would predict. Elsewhere, we show that voucher students in New York attended private schools that spent roughly half as much per pupil as did public school students (Howell & Peterson, 2002, p. 92).

27. See Chapter 6 of Howell and Peterson (2002) for test-score results from Dayton, Ohio, and Washington, D.C.

REFERENCES

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. (2003). 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*(462), 299-311.

Black, S. E. (1999, May). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, *114*, 577-599.

Bostic, R. W. (1997, January). The role of race in mortgage lending: Revisiting the Boston Fed study (Finance and Economics Discussion Series, No. 1997-2). Washington, DC: Board of Governors of the Federal Reserve System.

Bostic, R. W., & Surette, B. J. (2000, April). *Have the doors opened wider? Trends in homeownership rates by race and income* (Working paper). Washington, DC: Federal Reserve Board.

Bradbury, K. L., Mayer, C. J., & Case, K. E. (2001, May). Property tax limits, local fiscal behavior, and property values: Evidence from Massachusetts under Proposition 2(1)/(2). *Journal of Public Economics*, *80*, 287-311.

Bullard, R. D. (1994). Race and housing in a "New South" city: Houston. In R. D. Bullard, J. E. Grigsby, III, & C. Lee (Eds.), *Residential Apartheid: The American legacy.* Berkeley: University of California Press.

Bullard, R. D., Grigsby, J. E., & Lee, C. (1994). *Residential Apartheid: The American legacy.* Berkeley: University of California Press.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Reprint from *Handbook of research on teaching*. Boston: Houghton-Mifflin.

Chubb, J. E., & Moe, T. M. (1990). *Politics, markets, and America's schools*. Washington, DC: Brookings Institution.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High school achievement*. New York: Basic Books.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Corporation for the Advancement of Policy Evaluation with Mathematica Policy Research. (1996). *Evaluation of the New York City scholarship program*. Proposal submitted to Smith Richardson Foundation, December 11.

Davern, M. E., & Fisher, P. J. (2001). Household net worth and asset ownership: Household economic studies. In *Current population reports: The survey of income and program participation*. Washington, DC: U.S. Department of Commerce, Economics and Statistics Administration.

Evans, W. N., & Schwab, R. M. (1993). *Who benefits from private education? Evidence from quantile regressions*. College Park: Department of Economics, University of Maryland.

Figlio, D. N., & Stone, J. A. (1999). Are private schools really better? *Research in Labor Economics*, *1*(18), 115-140.

Frankenberg, E., & Lee, C. (2002, August). *Race in American public schools: Rapidly resegregating school districts*. Cambridge, MA: The Civil Rights Project, Harvard University Press.

Gerber, A., and Green, D. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, *94*(3), 653-664.

Goldberger, A. S., & Cain, G. G. (1982, April-July). The causal analysis of cognitive outcomes in the Coleman, Hoffer, and Kilgore report. *Sociology of Education*, *55*, 103-122.

Greene, J. P., Peterson, P. E., & Du, J. (1998). School choice in Milwaukee: A randomized experiment. In P. E. Peterson & B. C. Hassel (Eds.), *Learning from school choice* (pp. 335-356). Washington, DC: Brookings Institution.

Grogger, J., & Neal, D. (2000). Further evidence on the effects of Catholic secondary schooling. In *Brookings-Wharton papers on urban affairs: 2000.* Washington, DC: Brookings Institution.

Hayes, K. J., & Taylor, L. L. (1996, February). Neighborhood school characteristics: What signals quality to homebuyers? *Federal Reserve Bank of Dallas- Economic Review*, *4*, 2-9.

Hill, J., Rubin, D. B., & 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., & Peterson, P. (with Wolf, P., & Campbell, D.). (2002). *The education gap: Vouchers and urban schools*. Washington, DC: Brookings Institution.

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

James, F. J. (1994). Minority suburbanization in Denver. In R. Bullard, J. E. Grigsby, III, & C. Lee (Eds.), *Residential Apartheid: The American legacy.* Berkeley: University of California Press.

Jencks, C. (1985). How much do high school students learn? *Sociology of Education*, *58*, 128-135.

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

Munnell, A. H., Browne, L. E., McEneaney, J., & Tootell, G. M. B. (1992, October). *Mortgage lending in Boston: Interpreting HMDA data* (Working Paper WP-92-7). Boston: Federal Reserve Bank of Boston.

Munnell, A. H., Tootell, G. M. B., Browne, L. E., & McEneaney, J. (1996, March). Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review*, *86*(1), 25-53.

Neal, D. (1997). The effects of Catholic secondary schooling on educational achievement. *Journal of Labor Economics*, *15*(1), 98-123.

Peterson, P., & Howell, W. (2003). *Latest results from the New York City voucher experiment* (occasional paper 03-14, Program on Education Policy and Governance, Kennedy School of Government, Harvard University). Available at www.ksg.harvard.edu/pepg/

Peterson, P., & Howell, W. (2004). Efficiency, bias, and classification schemes: A Response to Alan B. Krueger and Pei Zhu. *American Behavioral Scientist*, *47*, 699-717.

Peterson, P. E., Howell, W. G., Wolf, P. J., & Campbell, D. E. (2003). School vouchers: Results from randomized experiments (pp. 107-144). In C. M. Hoxby (Ed.), *The economics of school choice*. Chicago: University of Chicago Press.

Peterson, P., Myers, D., Howell, W., & Kim, J. (1999). *An evaluation of the New York City school choice scholarships program: The first year.* Cambridge, MA: Program on Education Policy and Governance, Harvard University.

Popper, K. (1999). The logic of scientific discovery. In R. Boyd, P. Gasper, & J. Trout (Eds.), *The philosophy of science*. Cambridge, MA: MIT Press.

Rouse, C. E. (2000). *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 No. 440). Princeton, NJ: Princeton University Press.

Simmons, P. A. (1997). *Housing statistics of the United States* (1st ed.). Lanham, MD: Bernan.

U.S. Bureau of the Census. (2000). *The statistical abstract of the United States*. Washington, DC: Author.

U.S. Department of Commerce, Economics and Statistics Administration. (1999). *American housing survey for the United States*. Washington, DC: U.S. Census Bureau.

Weimer, D., & Wolkoff, M. (2001). School performance and housing values: Using non-contiguous district and incorporation boundaries to identify school effects. *National Tax Journal*, *LIV*, 231-253.

***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).*

***PAUL E. PETERSON** is a Henry Lee Shattuck Professor of Government at Harvard University. He 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 *(Brookings Institution Press, 2002).*