

Looking Inside the Black Box: What School Factors Explain Voucher Gains in Washington, DC?

Patrick J. Wolf and Daniel S. Hoople

*Public Policy Institute
Georgetown University*

Several evaluations of private school voucher programs in the United States have reported achievement gains for voucher users, particularly African Americans. These studies tend to be structured as Randomized Field Trials (RFTs) in which participants are assigned to treatment (offered a voucher) and control (not offered a voucher) groups by lottery. A major advantage of RFTs is that the randomization process controls for a number of factors, measurable and unmeasurable, that otherwise might confound the assessment of voucher effects. A major shortcoming of RFTs is that they tend to be “black box” evaluations that tell policy analysts precious little about why or how a policy intervention yields benefits downstream. In this article we use data from the 2nd-year RFT of the Washington, DC, privately funded voucher program, supplemented by information obtained from the schools that participating students attended, in an effort to identify what school features or practices might be boosting the achievement of voucher students. This preliminary analysis suggests that less extensive school facilities and programs, more advantaged peers, responsible teachers, and more time-consuming

We gratefully acknowledge the research assistance of Joshua Cowen, Jonathan Dreyfus, Brian Harrigan, and Juanita Riano, and the constructive comments of William Gormley, Jr., William Howell, Mihriye Mete, and Paul Peterson on earlier drafts of this article. We own any remaining errors.

Correspondence should be sent to Patrick J. Wolf, GPPI, The Car Barn, 4th Floor, 3520 Prospect Street, NW, Washington, DC 20007. E-mail: wolfp@georgetown.edu

homework assignments may be the characteristics of schools that increase the academic achievement of inner-city school voucher recipients. Additional analyses are recommended before definitive conclusions are drawn regarding what happens inside the black box of school voucher experiments.

Recently a number of major studies have been conducted on the effects of school vouchers on the educational achievement of students (Greene, 2001; Howell, Wolf, Campbell, & Peterson, 2002; Peterson, Myers, Howell, & Mayer, 1999). These studies tended to identify positive academic effects of vouchers on the African American students who used them to switch from public to private schools. The studies are designed as Randomized Field Trials (RFTs) with families assigned to treatment and control groups by lottery. The randomization process tends to control for all other factors, aside from the treatment and mere chance, that otherwise might confound the analysis. Therefore, analysts can ascribe to the treatment responsibility for any postintervention differences between the students who received vouchers and the control group who did not, with known levels of confidence (Boruch, de Moya, & Snyder, 2002, p. 51). Economist Carolyn Hoxby (2000a) described this methodological approach as the "gold standard" for research on educational interventions, such as vouchers, that otherwise would be vulnerable to self-selection bias.

Despite the advantages of these RFTs for correctly identifying the independent effects of school voucher interventions, they have at least one significant shortcoming: By themselves they do not indicate why vouchers improve the educational performance of the students who use them. RFTs tend to operate as classic "black box" analyses in which participants are baseline tested; one randomly determined group is permitted to receive treatment (enter the black box), whereas the control group does not, and the groups are retested after a certain period of time. What happens inside the box to generate posttest differences between the two groups is not the focus of most such studies. As Hess and Loveless (2005) stated, this "makes it difficult to know whether positive results can be replicated, negative results can be remedied, or what kinds of choice-based reforms are likely to do the most good for the largest number of children" (p. 1).

In contrast, a number of education studies have eschewed the strictures of random assignment and charged directly into the black boxes of effective schools (Carter, 2000; Kannapel & Clements, 2005; Thernstrom & Thernstrom, 2003). Such studies richly document the educational environment of high-performing schools serving disadvantaged students. However, because they lack an appropriate counterfactual, they cannot identify with certainty which common characteristics are actually causing effective performance.

School choice studies that explore the black box of schooling factors while retaining the critical methodological advantages of experimental design are needed. This article describes an initial foray into that largely unexplored wilderness. Using data from the 2nd-year evaluation of the Washington Scholarship Fund (WSF) school voucher program in 2000, we attempted to identify particular characteristics of schools that represent proximate causes for the greater academic achievement of voucher students. Our analysis is essentially exploratory and our findings are highly preliminary. They suggest that there is no single silver bullet that individually explains voucher gains. However, it appears that greater school resources, smaller schools, smaller class sizes, strict order and behavior, and increased communication between schools and parents do not explain the voucher gains, although such factors are often associated with successful schools. The best candidates for factors that are driving the educational improvement of voucher users are more racially diverse peers, more homework, and teachers exhibiting certain positive behaviors. More research will be necessary to confirm these initial findings.

Theory

The established theories regarding which school characteristics improve the performance of low-income, inner-city students are both rich and varied. They include claims that school resources; the social composition of schools, size, community, order, and discipline; and high expectations are the reasons why some disadvantaged urban students achieve more than others do. We consider each of these theories in turn.

Resources

Entire rain forests have been felled in the debate over whether increased resources measurably contribute to positive educational outcomes. Politicians certainly behave as if they believe that more resources purchase better education, as per-pupil spending on education has more than quintupled, in real terms, over the past 50 years (Brandl, 1998, pp. 2–3). Moreover, educational equity in the United States has often been defined in terms of the equalization of spending on students across local school districts and states, with courts intervening to order such equalization (Reed, 2001). Eric Hanushek (1996, 1989), among others (e.g., Brandl, 1998), argued repeatedly that additional investments of resources in K–12 education, in general, fail to produce consistent payoffs in the form of higher student achievement. Other scholars have adopted the moderate position that

more money improves educational outcomes, even dramatically, when it is invested properly (Hedges & Greenwald, 1996, Hedges, Laine, & Greenwald, 1994). Such a claim could be fully consistent with Hanushek's findings, if it is the case that most schools do not invest additional resources in the sorts of products and activities that yield positive achievement benefits for students. The investment problem might be particularly acute for inner-city schools that tend to serve those eligible for vouchers, as they may face more claims for, less discretion regarding, and less expertise in the effective deployment of resources.

Peers

A rich theoretical and empirical literature has established that characteristics of a student's peer group have independent and significant effects on the student's academic achievement (Dronkers & Robert, 2003; Hess & Leal, 1997; Nielsen & Wolf, 2001). Since the seminal Supreme Court ruling of *Brown v. Topeka Board of Education*, our society has operated under the maxim that an educational system that is entirely separated by race is inherently unequal. Thus, racial desegregation has long been thought to be an important vehicle for improving educational outcomes for low-income African American students. All other things being equal, we would expect disadvantaged minority students to learn more if they were to be educated in a racially diverse environment containing at least a modicum of White peers. Similarly, we also would expect lower income students to perform better academically if they were educated with higher income students as their peers (Benveniste, Carnoy, & Rothstein, 2002; Hoxby, 2002b).

Scale

Throughout the 20th century, a dominant maxim regarding K-12 education was that bigger is better. Harvard University President Joseph Conant famously argued that large elementary and secondary schools could take advantage of economies of scale to provide better facilities and more diverse academic programs and extracurricular activities to their larger student bodies (Ravitch, 2000, p. 363). However, recently there has been a movement toward a smaller-is-better mentality. Smaller schools, it is thought, are less imposing to students and better able to provide them with individual attention (Nathan & Febey, 2001). Smallness appears to be most important in the classroom, as two influential studies of a class-size reduction experiment in Tennessee determined that smaller classes enhance student performance—at least in the younger grades (Krueger, 1999; Mosteller, 1995).

Community

Schools serve disadvantaged students best when they function not as educational bureaucracies or businesses but as educational communities. Noted education analysts Brandl (1998); Bryk, Lee, and Holland (1993); and Coleman and Hoffer (1987) argued that students, especially those from disadvantaged backgrounds, thrive in private schools because of social capital that is produced by the nurturing of a strong sense of community. Private schools communicate more with parents and draw them into the school to participate, in a partnership, in the education of their child. Teachers have high morale and a strong sense of mission that is driven by concerns for the well-being of each student (Dronkers & Robert, 2003). They argue it is precisely this fostering of education as a communal enterprise that results in disadvantaged students learning more in private schools.

Strict Order and Behavior

Disadvantaged inner-city students are more likely to face disorder on a daily basis—in their neighborhoods, schools, and homes. The danger and unpredictability of chaotic environments naturally engenders fear and hopelessness in young people. Schools that are able to establish a safe and well-ordered environment, through strict rules and careful monitoring of the school and classrooms, likely serve as a firebreak to hopelessness and fear for low-income urban children. Moreover, disorder in the classroom distracts teachers from teaching and students from learning. It is generally recognized that an orderly learning environment is a necessary precondition for effective learning in school (e.g., Dronkers & Robert, 2003; Wilson, 1989, pp. 21–23).

Homework

A number of educational studies have demonstrated that students tend to live up to the high or low expectations that are set for them by parents and teachers (Akerlof & Kranton, 2002). Ferguson (1998) argued that expectations are at least partly to blame for the infamous Black–White test score gap of nearly 2 years in the achievement of comparable Black and White students by the time that they are seniors in high school. Less is expected of disadvantaged students in terms of educational achievement, and they tend to deliver only what is expected of them. If disadvantaged students are assigned challenging homework and high expectations for success in completing it, progress might be made in closing the test score gap.

All of the previous theories are possible candidates for explaining what happens inside the black box to produce voucher gains. However, even this extended list of theories fails to entirely exhaust the possible explanations. Another important peer group factor, the average academic ability of the entire student body, could not be measured consistently for this study, because different schools use different tests and performance standards. The leadership of school principals and the quality of teachers may also be driving voucher gains. Because we were unable to obtain direct measures of those and other classroom factors, we may not be able to identify all the proximate causes behind voucher gains. Still, the analysis of the data we present should shed some light on the validity of each of these competing explanations.

Data and Methods

Data

The core of the data that we use in this study comes from the 2nd-year evaluation of the WSF privately funded voucher program. WSF provides partial tuition scholarships of up to \$2,200 to families in the District of Columbia with household income at or below 270% of the federal poverty line. Families with income below the poverty line are eligible for the maximum scholarship amount, whereas families at 270% of poverty are eligible for about half of the maximum. Similar to vouchers, the scholarships can be redeemed at any of the more than 100 Washington, DC, and surrounding area private schools that participate in the scholarship program. WSF has been awarding scholarships to Washington, DC, students since 1993. In 2002, 1,325 elementary and secondary school students received scholarships.

The WSF experienced a dramatic expansion of their scholarship program in 1998. Because demand greatly exceeded supply (even after the addition of 1,000 new scholarships), the vouchers were awarded by lottery. Because only the luck of the draw determined which family would receive a voucher, the effect of the voucher on student and family outcomes could be studied via an RFT. In the spring of 1998, before the scholarships were awarded, the families of 1,582 students were surveyed about their family characteristics and school experiences. The students, who all were enrolled in Grades 1 through 8 of a public school at the time, completed the Iowa Test of Basic Skills (ITBS) in reading and math to produce a baseline measure of their academic abilities. After baseline data collection was com-

plete, 811 of the students in the study population were awarded scholarships by lottery. The remaining 771 students composed the control group for the study.

Because families in the WSF evaluation were assigned to the scholarship treatment or control groups by lottery, there was no need for the original evaluation—or this secondary study—to control for educationally relevant student and family traits such as income or race (Peterson & Howell, 2004). The treatment and control groups in the Washington, DC, evaluation did not differ significantly on any of nearly 50 individual student and parent characteristics (Wolf, Howell, & Peterson, 2000). The lottery worked.

Two years after the scholarship offer, in the spring of 2000, the remaining members of the treatment and control groups were invited to data collection sessions in which their children were retested, and the parents and older students were again surveyed about their educational experiences. Because 125 members of the control group had won the turnout incentive lottery in the first retest year of 1999, 1,457 students remained in the study population. The 2nd-year turnout of 730 students composed 50% of that population, for both the treatment and control groups.¹

The 2nd-year WSF evaluation data were generated via a number of measurement instruments. The baseline and 2nd-year test scores were generated by the performance of students on the ITBS in 1998 and 2000, respectively. Parental reports of school factors such as indicators of school community and order were obtained from survey instruments. The students in Grades 4 through 9 were surveyed about teacher attitudes, homework, and the level of strict discipline at their schools (Wolf, Peterson, & West, 2001).

These core data from the RFT were then supplemented by information collected from and about the various public and private schools that the students attended during the 1999–2000 academic year. The supplemental data included information about per-pupil spending on students as well as statistics regarding student body characteristics, enrollments, and class sizes. For the noncharter Washington, DC, public schools in the sample, these data were provided to us by the District’s Office of Public Accountability. For the private schools in the sample, the data were obtained from two sources. Information about the Catholic parochial schools was ob-

¹Because patterns of nonresponse to outcome data collection are probably nonrandom, individual observations were weighted to “rebalance” the sample so that the treatment and control groups remained similar regarding their baseline characteristics. For a complete description of this process, see Howell and Peterson (2002, pp. 209–216).

tained from the Office of the Superintendent of Schools for the Catholic Archdiocese of Washington, DC. Statistics regarding the nonparochial religious and independent private schools were obtained from responses to a mail survey. The response to the mail survey was high, thanks in part to our persistence in following up with private school administrators, as we obtained at least some information from over 80% of the private schools in the sample.

Still, moderate data gaps remain, especially regarding potentially important factors such as school size, class size, and the racial and income demographics of the schools. Where data regarding these four factors were initially missing for a particular private school, we replaced the missing data with the 2000 figures obtained by the Private School Survey (PSS; Broughman & Colaciello, 2001). For public schools initially missing responses for these variables, we imputed 2001–02 data from the Common Core of Data (CCD; National Center for Education Statistics, 2001–02), under the assumption that the size and social composition of a given school tends to change little over 2 years.² To prevent nonrandom list-wise deletion of observations with some remaining missing data, we replaced missing data in our explanatory variables with values of zero and included missing data dummy variables. By replacing the missing values, we avoided the inefficiency and potential bias caused by the exclusion of a significant proportion of the observations. By including a missing data dummy variable for any variable so altered we prevented the replacement zeroes from biasing the estimation of the beta coefficients in the statistical model (Cohen & Cohen, 1983, pp. 275–300). Still, including a significant number of such nontheoretical covariates reduces the efficiency of the estimations somewhat (i.e., expending one degree of freedom per missing data variable) and potentially biases estimates if the missing data dummies co-vary significantly with theoretically based explanatory variables (Peterson & Howell, 2004). To minimize the likelihood of such problems, each theoretical claim regarding the proximate cause of scholarship gains was tested using the two or three variables that represented the best indicators available of the influence of that particular factor. In the future, more data-rich, inside-the-black-box analyses might employ more comprehensive models of possible explanatory factors.

The Appendix presents descriptive statistics regarding the variables used in the analysis. Where the operational measurement of the variable is

²A total of 322 missing data elements were imputed this way, composing just 14% of the data for those four variables. Dummy variables signifying whether data were obtained using the PSS or CCD were also included in the analysis to capture any measurement error between the different sources.

not obvious, a description is provided in the notes. As revealed by the numbers in the far right column, the most significant missing data problems come from measures such as the Teacher Behavior Index and Strict Rules, which were drawn from student survey responses. That is because the third graders in the study were too young to be surveyed, and some of the students surveyed left questions blank. Several of the variables that rely on data drawn from the schools themselves also suffer from significant data gaps, especially those for student body demographics. Still, a reasonably large pool of actual data is available to estimate the relations hypothesized in this article.

Methods

Next we report on the effects of a private school scholarship offer on student test scores, which are also known as the intention-to-treat (ITT) effect. The ITT effect likely underestimates the effect of actually using a voucher to attend a private school, because some of the participants who are randomly assigned to the treatment fail to use it (i.e., they remain in public school) and some participants who are randomly denied the treatment obtain it anyway (i.e., enroll in private school without the aid of a voucher). In the analysis data, 49% of the students originally offered a scholarship were using it to attend private school in 2000, whereas 10% of the students in the control group were attending private schools using resources outside of the program.

We use the ITT effect as the focus of our study, even in the face of these high levels of “noncompliance” with the original assignment to the treatment condition for two reasons: (a) The original treatment assignment is the only grouping of students that is purely random in the sense that it is unaffected by subsequent participant behavior, and (b) even students who declined to use an offered scholarship may have used the inspiration or leverage from such an offer to secure a place in a preferred educational environment. Because we are seeking to identify the specific characteristics of schools that may be producing scholarship gains, it is both methodologically conservative and transparent to use a straightforward assignment-to-treatment variable as the source of the effect that needs to be “explained away” by factors causally downstream of the randomized scholarship offer.

Our attempt to account for the Washington, DC, scholarship gains proceeds in three steps. First, we conduct a correlation analysis to determine whether our schooling factor variables are strong candidates to explain the scholarship impact. The variables with the best chance of explaining observed voucher impacts on test-score outcomes would be those factors that

are significantly correlated with the treatment condition and also correlated with 2nd-year test scores (Howell & Peterson, 2002, pp. 158–159). By definition, any inside-the-black-box factor that explains an experimental impact must influence the outcome being studied and occur in different amounts or intensity for the treatment group.

Second, we estimate the general experimental impact of the programmatic treatment on 2nd-year test-score outcomes using three simple Ordinary Least Squares regression models:

$$Y_{2R} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \mu \quad (1)$$

$$Y_{2M} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \mu \quad (2)$$

$$Y_{2C} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \mu \quad (3)$$

In Model 1, Y_{2R} is each student's reading achievement score after 2 years on the ITBS expressed in National Percentile Rank (NPR) points.³ T is an indicator variable for whether a student was assigned randomly to the treatment group. Y_{0R} and Y_{0M} are the baseline reading and math scores. Baseline test scores are included as control variables to adjust for minor differences between the treatment and control groups on achievement on the baseline tests and to increase the precision of the estimated impacts. The β_1 coefficient therefore represents the estimated impact of the offer of a voucher on subsequent combined student test scores.

Equations 2 and 3 are variations on the basic regression model. They vary only in the measurement of the dependent variable. Equation 2 uses each student's math score as the achievement indicator. Equation 3 averages the reading and math scores into a composite measure of achievement. Reading and math scores are first estimated separately because reading is thought to be a more culturally affected outcome and math a more cognitive outcome. A particular schooling factor may not have the same influence on the test scores in both subject areas (Dronkers & Roberts, 2003, p. 15). The third equation consolidates the two, because composite test-score measures tend to provide more stable regression estimates (Howell & Peterson, 2002; Krueger, 1999).

The regression coefficients on the programmatic treatment variable in these three initial equations then become the benchmarks with which we compare the treatment coefficients in subsequent models that include variables representing the potential proximate causes of the voucher impact.

³For ease of interpretation, we report impacts in terms of NPR points. The results do not change substantively when using National Curve Equivalents or raw test scores.

The third step involves the estimation of the inside-the-black-box Ordinary Least Squares regression models. We estimate the following three slightly different models:

$$Y_{2R} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \beta_4 S + \mu \quad (4)$$

$$Y_{2M} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \beta_4 S + \mu \quad (5)$$

$$Y_{2C} = \alpha + \beta_1 T + \beta_2 Y_{0R} + \beta_3 Y_{0M} + \beta_4 S + \mu \quad (6)$$

These models are merely elaborations of Models 1 to 3. Models 4 to 6 add element S , which represents the vector of schooling factors hypothesized to affect test scores. The β_4 coefficients are the inside-the-black-box effects that are of central interest to the analysis. Should any one (or any combination) of these factors explain treatment students' test-score gains, their coefficients should be statistically significant and their inclusion in the models should substantially attenuate the observed scholarship effect.

We acknowledge here that ours is not the first attempt to “explain away” the 2nd-year voucher impacts in the WSF study. Howell and Peterson (2002, pp. 158–163), in their comprehensive report on experimental voucher impacts in New York City; Dayton, Ohio; and Washington, DC, peered inside their black boxes to try to identify the proximate causes of the effects. They included parental reports of school resources, school programs, disruptions, communication, school size, and class size in their regression models. Although school disruptions, school size, and class size were found to be associated with 2nd-year, test-score outcomes in Washington, DC, their measure of the voucher impact on test scores actually increased when those schooling factors were included in the model. Howell and Peterson concluded that “all the elements that parents identified as distinctive about private schools ... could not explain the private school advantage for African American students” (pp. 162–163).

Our study differs from the previous Howell and Peterson (2002) inside-the-black-box analysis of the 2nd-year Washington, DC, voucher impacts in ways that we think increase the likelihood of explaining the scholarship impact. First, we use the entire set of 2nd-year respondents in our analysis—not just the subset of African American respondents used by Howell and Peterson. Although this adds only 36 observations to the analysis, for our purposes there is no reason to exclude the 36 non-Black study participants.⁴ Second, we use the scholarship offer as the treatment impact

⁴Howell and Peterson (2002) restricted their inside-the-black-box analysis to African Americans because their consistent finding across cities was that the vouchers boosted the test scores only of African Americans.

to be explained away, whereas Howell and Peterson focused on the effect of private schooling as their treatment variable. Third, and most important, we add additional school-level variables to the analysis, drawn from administrative data provided by the schools. In thereby bringing more evidence to the table, we expect to have a better chance of identifying the proximate causes of the previously reported Washington, DC, voucher gains.

We also acknowledge that the 2nd-year Washington, DC, voucher impacts that we analyze here were not stable over time. The 1st-year experimental test-score impacts varied by subject matter and grade cohort, with younger students exhibiting voucher gains in math but older students signaling voucher losses in reading (Wolf et al., 2000). The 2nd-year impacts were the most consistent, with both grade cohorts of students demonstrating voucher gains in both subject areas (Wolf et al., 2001). In the 3rd year of the evaluation, the WSF program experienced high levels of program attrition, and no experimental test-score impacts were detected (Howell & Peterson, 2002, pp. 145–150). We focus on the 2nd-year voucher impacts not to mislead readers about the ultimate success of the intervention, but because our goal is to try to explain why and how voucher programs boost achievement. In doing so, it makes sense to attempt to explain away voucher gains where and when they exist.

Results

Our correlation analysis indicates that the educational environment differed in important ways for the group of students offered vouchers (Table 1). The offer of a voucher was significantly associated with lower per-pupil expenditures, smaller schools, fewer peers receiving free or reduced price lunches, less extensive school facilities and programs, more homework, and more extensive school-home communication.⁵ Fewer peers receiving free or reduced-price lunch and more homework were also significantly correlated with outcome test scores, making those two school factors prime candidates for the proximate causes of any observed voucher impacts. Nevertheless, all 12 potential explanatory variables were significantly correlated with either the treatment condition or test-score out-

⁵Surprisingly, three characteristics often associated with private schools—caring and demanding teachers, White students, and an ordered school environment—were completely uncorrelated with the treatment condition. This is a result that could be peculiar to Washington, DC.

Table 1

How Correlated Are School Characteristics With Scholarship Offer and Test Scores?

Variables	Offered Scholarship		Combined Reading and Math Score in 2000		Valid N
	ρ	<i>p</i>	ρ	<i>p</i>	
Explanatory					
Per-pupil expenditure	-.37	< .01	-.03	.43	697
School enrollment	-.31	< .01	.02	.60	630
Free/Reduced lunch	-.17	< .01	-.20	< .01	481
School Facility Index	-.16	< .01	-.05	.23	704
Student estimated homework	.12	< .01	.13	< .01	527
Communication Index	.09	.02	.04	.24	698
Strict rules	.07	.15	.11	.02	459
Parent estimated homework	.06	.09	.09	.01	693
Class size	.06	.15	.16	< .01	573
Teacher Behavior Index	-.02	.61	.17	< .01	439
White students	.01	.73	.18	< .01	578
School Order Index	.00	.96	.11	< .01	730
Control					
Baseline reading score	.00	.99	.41	< .01	730
Baseline math score	.00	.91	.58	< .01	730

Note. Correlations and significance levels derived using a pairwise correlation matrix. ρ = Pearson's bivariate correlation coefficient. Observations weighted to correct for nonresponse.

comes (or both), so we decided to include all of them in the subsequent inside-the-black-box estimation models.

Of importance, the control variables for baseline (i.e., prerandomization) test scores were completely uncorrelated with the treatment condition. This result provides additional confirmation that the scholarship lottery was truly random and the experimental groups were similar, on average, regarding individual baseline characteristics that tend to confound many non-experimental analyses of educational impacts (Cook & Payne, 2002, p. 173).

The results of our search for causal factors behind the scholarship gains appear in Table 2. The first row under the column headers in each table simply restates the average impact of the scholarship offer drawn from the Base Models 1 to 3, absent any of the schooling explanatory variables. The next row of results describes the residual scholarship offer impact after the effects of schooling factors in the model are accounted for. Should that figure become indistinguishable from zero due to the significant test-score influences of one or more explanatory variables in the model, then we have uncovered a plausible inside-the-black-box explanation for the scholarship impact.

Table 2
Do School Characteristics Explain Scholarship Gains?

Factor	Reading Score (Model 1 or 4)		Math Score (Model 2 or 5)		Combined Score (Model 3 or 6)	
	β	SE	β	SE	β	SE
Base model (1, 2, 3)	2.69**	1.35	3.67***	1.35	3.18***	1.13
Scholarship effect alone						
Expanded model (4, 5, 6)						
Scholarship effect	1.92	1.44	3.79***	1.43	2.85**	1.18
Explanatory variables						
White students	.16***	.06	.03	.06	.09*	.05
Per-pupil expenditure (in thousands)	-.31	.49	.03	.48	-.14	.40
Free/Reduced lunch	.02	.04	-.03	.04	-.004	.03
School enrollment (in hundreds)	.56	.39	1.10***	.38	.83***	.32
Class size	.25*	.15	.12	.15	.18	.12
School Facility Index	-.58**	.24	.09	.24	-.25	.20
Teacher Behavior Index	.87***	.31	.52*	.31	.70***	.25
Communication Index	.15	.41	.21	.41	.18	.34
School Order Index	.15	.14	.10	.14	.12	.11
Strict rules	.96	.90	.19	.90	.58	.74
Parent estimated homework (hr/day)	.52	1.24	.08	1.23	.30	1.01
Student estimated homework (hr/day)	.06	1.08	2.06*	1.08	1.06	.88
Control variables						
Baseline reading score	.22***	.03	.07***	.03	.15***	.02
Baseline math score	.34***	.03	.41***	.03	.38***	.03
Adjusted R ²	.38	—	.34	—	.45	—
N	730	—	730	—	730	—

Note. Observations are of students in Grades 3 to 9 in 2000. Observations weighted to correct for nonresponse. Missing data and Common Core of Data dummy variables included. * $p < .10$. ** $p < .05$. *** $p < .01$ (two-tailed).

Only one of the initial explanatory models (Model 4) approximates our goal. The scholarship impact on reading scores drops by about 30%, from 2.7 NPR to 1.9 NPR, and loses statistical significance. Exposure to more White peers, larger class sizes, less extensive facilities, and positively behaved teachers significantly contribute to higher reading scores in Model 4. However, of those four schooling factors, only less extensive facilities demonstrated a significant bivariate correlation with the offer of a scholarship. Teachers exhibiting positive behaviors, larger schools, and more homework are associated with higher math achievement in Model 5; how-

ever, the independent impact of the scholarship offer remains statistically significant and similar in size to the original base estimate. By combining the reading and math scores into a single measure of achievement to be estimated, Model 6 essentially averages the effects of Models 4 and 5. More White peers, larger schools, and better behaved teachers explain significant levels of variation in combined test-score outcomes, but so too does the scholarship offer. The Model 4 to 6 estimations do not appear to provide a very convincing explanation for why the scholarship offer led to higher test scores after 2 years.

The factors that do not appear to explain voucher gains are also interesting. Consistent with Hanushek's (1996, 1989) research, per-pupil expenditures are not consistently associated with test-score outcomes in the sample. The percentage of peers enrolled in the free or reduced priced lunch program also did not systematically affect achievement, however measured. School-home communication, school order, and strict rules similarly did not significantly influence test-score gains. Finally, parent reports of homework burden did not predict test-score outcomes, even as student reports influenced math gains. Students may be providing the more accurate numbers here, as the homework burden falls more directly on them.⁶

Discussion

What are we to make of these results? First of all, there appears to be no smoking gun. Responsible teacher behavior was the only schooling factor that significantly explained test-score outcomes in all three variations of the inside-the-black-box models. Only the Model 4 estimation of reading outcomes rendered the scholarship effect indistinguishable from zero. In addition, the set of schooling factors associated with achievement gains in reading differed rather sharply from the set that explained gains in math. Apparently, when looking inside the black box for explanations of school voucher gains, we need to treat reading and math scores as separate outcomes.

Nevertheless, several schooling factors have emerged as possible explanations for the scholarship achievement gains in Washington, DC. Less extensive facilities and programs are more common in the schools attended by students offered scholarships, and that school characteristic is associated with subsequent achievement gains in reading. At first blush, that as-

⁶Although the parent and student estimates of homework burden were highly correlated, the results of the models do not change noticeably if either measure is estimated with the other omitted.

sociation may seem to be perverse. How could less extensive facilities and programs at a school enhance reading performance? There are two reasonable explanations. First, less extensive school facilities may be an indicator that a school is more focused on its core mission of advancing learning inside the classroom. Separate gymnasiums, lunchrooms, computer labs, and counselors' offices may serve various individual needs; however, they may also divert time and financial resources away from the sort of focused classroom instruction that seems to be central to reading mastery. Schools that eschew such amenities may be left with more resources to direct toward advancing learning. Second, fewer special programs for students may be an indicator of a school's commitment to educating all students together. Bryk et al. (1993) noted that Catholic schools, in particular, are less likely to direct particular students to special educational tracks or programs, and that their communal attitude of "everybody together" is central to their success as educational institutions.

Teacher behavior also emerged as a likely factor in explaining scholarship student reading gains in Washington, DC. Teachers who are described by students as interested in them, good listeners, fair, respectful, and willing to punish cheaters probably are more effective at motivating and teaching students to read. However, the bivariate correlations suggest that such teacher attributes were just as common in the educational environments of the control group students as the scholarship students. However, conditional on other schooling characteristics, positive teacher behaviors appear to contribute significantly to learning and explain at least some of the voucher impacts observed in Year 2 of the Washington, DC, experiment.

Ethnic diversity appears to be an additional factor that partly explains away the voucher gains in reading. The percentage of White students in the school influenced reading and combined achievement outcomes, and it appears to have contributed to the reduction in the voucher impact for reading. Unfortunately, White students are in limited supply in the district public and private schools attended by students in the study, as only 5% of their schooling peers were White. Racial diversity appears to benefit low-income Washington, DC, students but is difficult to achieve on a large scale, given the demographics of the schooling population. Larger class sizes also boosted reading test scores, perhaps because principals assign their most able teachers to their largest classes.

The amount of homework assigned stood out as a schooling factor that boosted math test-score achievement. As an inherently cognitive process, math skill acquisition may be especially sensitive to the amount and difficulty of the problem-solving work sent home with students on a regular basis. Because the bivariate correlations revealed that scholarship students

tended to receive more homework than their control group counterparts, the amount of homework has emerged as an important piece of the puzzle regarding scholarship gains in math in Washington, DC.

The remaining schooling factor that influenced math and combined achievement—higher school enrollments—may have seemed surprising. This result could be a reflection of a “flight to quality.” Especially in Washington, DC, where a high percentage of parents have the opportunity to place their child in a private, public, or charter school of their choosing, good schools might quickly become large. Teachers trained in math instruction are in short supply in education. Larger schools likely have a deeper pool of teachers, increasing the likelihood that more of them are trained specifically to teach math.

Conclusion

Just as Rome was not built in a day, the answer to the vexing question of what school characteristics or practices might explain the gains evidenced by scholarship users is not likely to emerge, decisively, from a preliminary analysis of data from a single school voucher experiment. It is left to future analyses to generate more conclusive results regarding the proximate causes of voucher test-score gains. Nevertheless, this initial foray inside the black box of school voucher impacts provides important lessons for future researchers. First, they should separately analyze the factors that influence reading and math achievement gains, as our results revealed significant differences in the characteristics associated with performance in those two subject areas. Second, the analysts who conduct such probes should especially seek to obtain more precise measures of the academic focus and overall educational quality of schools, as the facilities-program and school-size variables used in this analysis appeared to be acting as proxies for those largely unmeasured but important factors. The behavior of teachers and the amount of homework also have emerged from this study as possible components of a comprehensive explanation for scholarship student gains in Washington, DC, and the importance of those factors should be further explored. Racial diversity appears to play a role in explaining voucher impacts, to the extent it can be achieved in an urban educational environment such as Washington, DC.

For now, we should draw only tentative conclusions regarding what happens inside the black box to generate voucher test-score gains. We clearly need more analyses that peer into future voucher black boxes, we hope with a more powerful and carefully calibrated microscope.

References

- Akerlof, G. A., & Kranton, R. E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature*, 40, 1167–1201.
- Benveniste, L., Carnoy, M., & Rothstein, R. (2002). *All else equal*. New York: Routledge Falmer.
- Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50–79). Washington, DC: Brookings Institute.
- Brandl, J. E. (1998). *Money and good intentions are not enough*. Washington, DC: Brookings Institute.
- Broughman, S., & Colaciello, L. (2001). *Private School Universe Survey: 1999–2000* (NCES 2001330). Washington, DC: National Center for Educational Statistics, U.S. Department of Education. Available from <http://nces.ed.gov/pubs2001/2001330.pdf>
- Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.
- Carter, S. C. (2000). *No excuses: Lessons from 21 high-performing, high poverty schools*. Washington, DC: Heritage.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Coleman, J. S., & Hoffer, T. (1987). *Public and private high schools: The impact of communities*. New York: Basic Books.
- Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in education research. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 150–178). Washington, DC: Brookings Institute.
- Dronkers, J., & Robert, P. (2003). *The effectiveness of public and private schools from a comparative perspective* (European University Institute Working Paper, SPS No. 2003/13). Fiesoli, Italy: European University Institute.
- Ferguson, R. F. (1998). Teachers' perceptions and expectations and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test-score gap* (pp. 273–313). Washington, DC: Brookings Institute.
- Greene, J. P. (2001). Vouchers in Charlotte. *Education Matters*, 1(2), 55–60.
- Hanushek, E. A. (1996). School resources and student performance. In G. Burtless (Ed.), *Does money matter?* (pp. 43–73). Washington, DC: Brookings Institute.
- Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, 18, 45–51.
- Hedges, L. V., & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G. Burtless (Ed.), *Does money matter?* (pp. 74–92). Washington, DC: Brookings Institute.
- Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, 23(3), 5–14.
- Hess, F. M., & Leal, D. L. (1997). Minority teachers, minority students, and college matriculations: A new look at the role-modeling hypothesis. *Policy Studies Journal*, 25, 235–248.
- Hess, F. M., & Loveless, T. (2005). How school choice affects student achievement. In J. Betts & T. Loveless (Eds.), *Getting choice right: Ensuring equity and efficiency in education policy* (pp. 85–100). Washington, DC: Brookings Institute.
- Howell, W. G., & Peterson, P. E. (with Wolf, P. J., & Campbell, D. E.). (2002). *The education gap: Vouchers and urban schools*. Washington, DC: Brookings Institute.

- Howell, W. G., Wolf, P. J., Campbell, D. E., & 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.
- Hoxby, C. M. (2000a, March 9). Discussant commentary on the panel on voucher effects at the conference on “Vouchers, Charters, and Public Education,” Program on Education Policy and Governance, Harvard University, Cambridge, MA. (Transcript available upon request.)
- Hoxby, C. M. (2000b). *Peer effects in the classroom: Learning from gender and race variation* (National Bureau of Economic Research Working Paper 7867). Cambridge, MA: National Bureau of Economic Research.
- Kannapel, P. J., & Clements, S. K. (2005). *Inside the black box of high-performing high-poverty schools*. Report of the Prichard Committee for Academic Excellence. Lexington, KY: Prichard Committee for Academic Excellence. Available at <http://www.prichardcommittee.org/Ford%20study/FordReportJE.pdf>
- Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114, 497–532.
- Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 5, 113–127.
- Nathan, J., & Febey, K. (2001). *Smaller, safer, saner, successful schools*. Washington, DC: National Clearinghouse for Educational Facilities.
- National Center for Education Statistics. (2001–02). *Common core of data*. Washington, DC: Author. Available from <http://nces.ed.gov/ccd/>
- Nielsen, L. B., & Wolf, P. J. (2001). Representative bureaucracy and harder questions: A response to Meier, Wrinkle, and Polinard. *Journal of Politics*, 63(2), 598–615.
- Peterson, P. E., & Howell, W. G. (2004). Voucher research controversy: New looks at the New York City evaluation. *Education Next*, 4, 73–78.
- Peterson, P. E., Myers, D., Howell, W. G., & Mayer, D. (1999). School choice in New York City. In P. E. Peterson & S. Mayer (Eds.), *Earning and learning: How schools matter* (pp. 317–340). Washington, DC: Brookings Institute.
- Ravitch, D. (2000). *Left back: A century of failed school reforms*. New York: Simon & Schuster.
- Reed, D. S. (2001). *On equal terms: The constitutional politics of educational opportunity*. Princeton, NJ: Princeton University Press.
- Thernstrom, A., & Thernstrom, S. (2003). *No excuses: Closing the racial gap in learning*. New York: Simon & Schuster.
- Wilson, J. Q. (1989). *Bureaucracy*. New York: Basic Books.
- Wolf, P. J., Howell, W. G., & Peterson, P. E. (2000). *School choice in Washington, D.C.: An Evaluation after one year* (Program on Education Policy and Governance Working Paper PEPG/00–08). Cambridge, MA: Harvard University Press.
- Wolf, P. J., Peterson, P. E., & West, M. R. (2001). *Results of a school voucher experiment: The case of Washington, D.C. after two years* (Program on Education Policy and Governance Working Paper PEPG/01–05). Cambridge, MA: Harvard University Press.

Appendix

Table A1

Descriptive Statistics of Variables for Washington, DC, Study Participants in 2000

<i>Variables</i>	<i>M</i>	<i>SD</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Valid N</i>
Dependent					
Combined reading and math score in 2000	22.9 NPR	19.3 NPR	0 NPR	93 NPR	730
Reading score in 2000	26.2 NPR	22.2 NPR	0 NPR	99 NPR	730
Math score in 2000	19.7 NPR	21.4 NPR	0 NPR	95 NPR	730
Explanatory					
Scholarship award	54.7%	49.8%	0	1	730
Private school	31.4%	46.4%	0	1	730
Elite private school ^a	2.5%	15.6%	0	1	726
Per-pupil expenditure ^b	\$5,824	\$2,071	\$395	\$18,000	697
Facility/Programs Index ^c	9.3	3.1	1	14	704
White students	5.0	15.7	0	98	578
Free/Reduced lunch	67.0	28.7	0	99	481
School enrollment	399.9	235.5	18	2,171	630
Class size	17.1	5.6	4	56.7	573
Communication Index ^d	5.9	1.8	1	8	698
Teacher Behavior Index ^e	15.7	2.8	6	20	439
School Order Index ^f	17.7	4.7	7	21	730
Strict rules ^g	2.9	.9	1	4	459
Parent estimated homework	74 min	35 min	0 min	150 min	693
Student estimated homework	69 min	45 min	15 min	165 min	527
Control					
Baseline reading score	30.3 NPR	27.1 NPR	0 NPR	99 NPR	730
Baseline math score	23.1 NPR	21.8 NPR	0 NPR	99 NPR	730

Note. Restricted to students in Grades 3 to 9. NPR = National Percentile Ranks.

^aDefined as private but unaffiliated with a religious organization. ^bDefined as the regular tuition level at private schools and the district per-pupil expenditure level (net of special education) for public schools, differentiated by neighborhood public and public charter. ^cBased on parental responses regarding the presence or absence of a computer lab, a library, a gym, a cafeteria, special programs for non-English speakers, individual tutors, special programs for slow learners, special programs for fast learners, child counselors, a nurse's office, a music program, an arts program, an after-school program, and prepared lunches. ^dBased on parental responses to questions regarding whether they are sent midterm grades, notified when student is disruptive, asked to talk to class about their job, asked to assist in instruction, asked to attend open houses at school, asked to attend parent-teacher conferences, sent regular notes from the teacher about student, and sent a school newsletter. ^eBased on student Likert-scale responses to whether they agree that their teachers are interested in students, really listen to them, are fair, avoid putting down students, and punish cheating when they observe it. ^fBased on parental reports reporting the relative seriousness (*very, somewhat, not*) of the following disorders: kids destroying property, tardiness, truancy, kids fighting, kids cheating, racial conflict, and guns or other weapons. ^gBased on student Likert-scale responses to whether "rules for behavior are strict."