Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning

Stephen L. Morgan
Cornell University

School-effects research in sociology cannot be separated from concerns about causality. Purely descriptive modeling justifications are untenable. Focusing on the Catholic school effect on learning, this article demonstrates an approach that places regression modeling strategies within a specific and well-developed framework for thinking about causality. While regression models should properly remain the workhorse methodology for school-effects research, regression estimates should more often be subject to exacting interpretations and presented alongside alternative estimates of more specific parameters of interest. In this demonstration, propensity-score matching estimates of the Catholic school effect for the Catholic schooled are provided to supplement the estimates obtained by regression models. Although subject to their own set of weaknesses, the matching estimates suggest that the Catholic school effect is the strongest among those Catholic school students who, according to their observed characteristics, are least likely to attend Catholic schools. Four alternative explanations are offered for this finding, each of which should be pursued in further research.

In a series of widely read research reports, Coleman and his colleagues (Coleman and Hoffer 1987; Coleman, Hoffer, and Kilgore 1982; Hoffer, Greeley, and Coleman 1985) presented evidence that Catholic schools confer learning advantages on their students. Although vigorously contested (see Alexander and Pallas 1983, 1985; Goldberger and Cain 1982; W. R. Morgan 1983; Noell 1982; Willms 1985), their provocative findings inspired three strands of subsequent survey research in the sociology and economics of education: the evaluation of market competition models of school improvement (e.g., Chubb and Moe 1990; Figlio and Stone 1997; Hoxby 1996; Neal 1997), effective schools research (e.g., Lee and Smith 1993, 1995; Lee, Smith, and Croninger 1997), and social capital research (e.g., Carbonaro 1998; S. L. Morgan and Sørensen 1999).

In most of this school-effects research (and, to be fair, in my research as well), the limitations of observational survey data are acknowledged but rarely discussed in any depth. Thus, although the possible existence of omitted variable bias is recognized, the care with which the weaknesses of survey data are discussed declines dramatically when the specter of hidden self-selection bias arises. An important negative consequence of suppressing forthright discussion of the specific weaknesses of available data and the limited range of conclusions that they can effectively sustain is that the need for more informative data to resolve empirical questions of theoretical and policy relevance is not clearly articulated.

Although the primary goal of this article is to demonstrate the utility for school-effects research of the careful application of a highly developed framework for thinking about causality and the evaluation of treatment effects, I also make a substantive contribution to the school-effects literature. A range of plausible estimates for the Catholic school effect on learning is needed. The estimates that currently inform policy debates on school choice and voucher programs (see Ladd 1996; Peterson and Hassel 1998; Rasell and Rothstein 1993) rely too heavily on a limited set of estimates from past research (most commonly, Chubb and Moe 1990). The propensity-score matching estimates presented in the last section of this article, although not without their own set of weaknesses, represent a set of plausible estimates that fill in the picture suggested by these controversial but commonly cited regression estimates.

Most important for policy, however, is that all these estimates—regression and propensity-score matching estimates—cannot tell us in any straightforward way how much public school students would benefit from attending Catholic schools. The counterfactual framework clearly demonstrates why this is the case and can be used explicitly to justify an appeal for the collection of more informative data.

In the remainder of the article, I first outline the basic regression strategy for the estimation of the causal effect of Catholic schooling on learning. With more recent data, I then estimate the regression models developed by Hoffer et al. (1985). After I introduce the counterfactual model of causality, I offer an alternative set of propensity-score matching estimates. In conclusion, I discuss implications of the results for policy debates on school choice and vouchers and for further research on school effects in sociology. Sandwiched between the empirical analyses, the main thrust of the article is the core set of ideas discussed in the application of the counterfactual model of causality, for these ideas suggest powerful conclusions on their own without reference to any specific estimates.

REGRESSION ESTIMATION OF THE CATHOLIC SCHOOL EFFECT

Do high school students who attend Catholic schools learn more than high school students who attend public schools? One way to answer to this question is to estimate a regression equation of the form:

$$(1) \quad Test12_i = a + d(Cath_i) + e_i$$

where i indexes sampled individuals and the dependent variable *Test*12 represents a score on a standardized cognitive test taken near the end of the senior year of high school. The sole independent variable, *Cath*, for Equation 1 is a dummy variable that equals 1 for students who attend Catholic schools and 0 for students who attend public schools. If *Test*12 adequately measures the cumulative amount of material learned by a student, an estimate of the intercept coefficient a is an estimate of the achievement of public school students, and an estimate of the coefficient d is an estimate of the difference in achievement between those who attend Catholic schools and those who attend public schools.

In most research in the sociology of education, it is assumed that learning is determined by family background characteristics through a collection of mediating mechanisms, only some of which are typically measured in large-scale surveys. If learning is some function of family background and if Catholic school students and public school students, on average, differ in their family background characteristics, then the achievement difference parameterized by the coefficient d in Equation 1 may not properly represent the Catholic school effect on learning.

The most common solution is to perform a covariance adjustment by estimating a regression equation of the form:

$$(2) \quad Test12_i = a + d(Cath_i) + b_i X_{1i} + \ldots + b_k X_{ki} + e_i$$

where X_1 through X_k are measured variables for family background characteristics. In almost all the research on the Catholic school effect on achievement, it has also been assumed that the variables X_1 through X_k should include more than just measures of family background characteristics (see Noell 1982:124 for the basic argument). The incorporation of additional variables in X_1 through X_k is straightforward although not without analytic cost, as is discussed later.[1]

One additional independent variable, however, has received special analytic attention. Consider the inclusion of a variable for prior test scores, as in a regression equation of the form:

$$(3)\quad Test12_i = a + d(Cath_i) + l(Test10_i) + b_1X_{1i} + \ldots + b_kX_{ki} + e_i$$

where the variable *Test*10 is a score from the administration of the same standardized test in the 10th grade. *Test*10 is best regarded as a value on the dependent variable from an earlier period. Except under extreme circumstances (e.g., when there is no measurement error in *Test*10 and all determinants of *Test*10 that differ across Catholic school students and public school students are included and properly specified as variables in X_1 through X_k), *Test*10 will be correlated with the error term e of Equation 3. As a result, estimates of the parameter l may be biased and inconsistent.

As has been shown in many methodological works (e.g., Allison 1990; Judd and Kenny 1981), bias in the lag coefficient l can induce substantial bias in estimates of other coefficients, such as d, when there are systematic differences across the two groups under comparison in the distribution of *Test*10. This result can be seen by subtracting $l(Test10)$ from both sides of Equation 3:

$$(3a)\quad Test12 - l(Test10_i) = a + d(Cath_i) + b_1X_{1i} + \ldots + b_kX_{ki} + e_i$$

If an estimate of l, $\hat{l}$, is attenuated by measurement error or otherwise biased toward zero, then *Test*12 will not be sufficiently adjusted for the true difference in prior achievements. This underadjustment induces a correlation between *Cath* and the error term, which upwardly biases an ordinary least-squares (OLS) regression estimate of d.[2] Although these potential problems are widely recognized, it is often difficult to know how severe the bias may be for any particular set of estimates.

In sociology, the most common representation of Equation 3 is produced by subtracting *Test*10 from both sides of Equation 3. The resulting lagged change-score model

$$(3b)\quad (Test12 - Test10)_i = a + d(Cath_i) + (l - 1)(Test10_i) + b_1X_{1i} + \ldots + b_kX_{ki} + e_i$$

yields the same estimate for the coefficient d as the covariance adjustment model written as either Equation 3 or Equation 3a. Equation 3b seems to be favored for presentation purposes because it is easier to motivate a model of learning with this specification (see S. L. Morgan and Sørensen 1999:Appendix B; Sørensen 1996).

Regression Estimates with NELS Data

In this section, I present regression estimates of the coefficient d from Equations 1, 2, and 3b. The three main goals of the analysis are to (1) demonstrate the regression approach to estimating the Catholic school effect based on the approach adopted by Coleman and his colleagues, (2) obtain regression estimates of the Catholic school effect with the best data currently available, and (3) set the stage for an interpretation of the regression estimates using the counterfactual model of causality.

To obtain the following estimates, I first extracted a set of variables and an analysis sample from the National Education Longitudinal Study (hereafter, NELS) that mirrors as closely as possible the variables and sample from the High School and Beyond (hereafter, HS&B) study that Coleman and his colleagues modeled in the best of their articles (Hoffer et al. 1985). Appendix A presents a description of the NELS data, a discussion of their comparability with the HS&B data that Hoffer et al. analyzed, basic summary statistics of the variables analyzed in this article, and an analysis of the consequences of different procedures for the handling of missing data in both Hoffer et al. and in the regression models I offer.

Table 1 presents regression estimates of the Catholic school effect for both mathematics and reading from models that employ best-subset regression imputation of item-specific missing values on the covariates X_1 through X_k. Model 1 estimates the simple dummy variable regression equation presented in Equation 1. The Catholic school coefficients listed in the first row of each panel correspond to estimates of the coefficient d in Equation 1 for the mathematics test and for the reading test, respectively.

With the senior-year mathematics test score as the outcome variable, the estimated coefficient for the Catholic school variable is 5.78, indicating that Catholic school students score almost six points higher on the test than do public school students. Although the magnitude of this effect is difficult to judge, it represents approximately 40 percent of the standard deviation of the senior-year test score, 13.96 (reported, along with other descriptive statistics in Appendix A, Table A1). Likewise, with the senior-year reading test score as the outcome variable, the estimated coefficient is 3.98, which is also approximately 40 percent of its corresponding standard deviation, 9.98. However, because measurement error inflates the variances of the test scores to an unknown degree, the true baseline Catholic school effect on achievement is most likely larger than 40 percent of the true standard deviation of senior-year achievement.

Model 2 introduces, as variables X_1 through X_k in Equation 2, the 15 variables listed under the subhead family background and other demographic characteristics in Table A1 (see Appendix A). The introduction of these variables into the estimated regression equation reduces the estimate of the Catholic school coefficient d by more than half for both the mathematics test and the reading test. The standard interpretation for this decline is that the learning that is produced by family background accounts for some of the baseline observed Catholic school effect on learning specified in Equation 1 and estimated as Model 1 in Table 1.

Model 3 adds to the independent variables of Model 2 the educational expectations and parental involvement variables presented in Table A1. The inclusion of these additional six variables further reduces the estimate of the Catholic school effect coefficient d. When it is recognized, in the language of Coleman et al. (1982:138) that these variables are "not clearly prior to the student's achievement," it is hard to know whether the estimated coefficient for d from Model 3 should be preferred to the estimated coefficient d from the simpler Model 2. If there is a true Catholic school effect on learning, most sociological theory would maintain that at least some portion of this effect would be produced by increases in students' own educational expectations, the expectations parents have of them, and the involvement of parents in the lives of students.

The first panel of Table 1 reports three more models that include additional variables in X_1 through X_k. Because I felt especially uncomfortable imputing values for the missing values on categorical course-taking variables, Model 4 simply reestimates Model 3 for the subset of respondents who do not have missing data on course-taking variables (e.g., for the mathematics tests, 10,003 of the 10,835 respondents for Model 3). The estimated coefficient for the Catholic school effect d is similar for Models 3 and 4, providing suggestive evidence that proceeding to further analysis with the narrowed sample does not introduce substantial bias into estimates of d.

Adding the school climate, curriculum track, and course-taking variables presented in Table A1 to the independent variables included in Model 4, Model 5 estimates an even more comprehensive regression equation. For Model 5, these additional variables reduce the Catholic school effect to a substantial negative value for the mathematics test and to a value close to zero for the reading test. Finally, adding a dummy variable for whether the student attended a Catholic middle school in the eighth grade to the independent variables included in Model 5, Model 6 estimates a Catholic school effect on senior-year achievement that is further reduced, becoming substantially negative for

Table 1. OLS Regression Estimates of the Catholic School Effect on Achievement

	Model: $Test12 = a + d(Cath_i) + b_1 X_{1i} + b_k X_{ki} + e_i$			Model: $(Test12 - Test10)_i = a + d(Cath_i) + (l - 1)(Test10_i) + b_1 X_{1i} + \ldots + b_k X_{ki} + e_i$		
	Regression Estimate of the Coefficient *d*	*SE*	*N*	Regression Estimate of the Coefficient *d*	*SE*	*N*
Mathematics Achievement Models						
1. Catholic school dummy variable only	5.78	.84	10,835	1.31	.31	10,835
2. 1 + Family background and demographics	2.54	.79	10,835	.99	.33	10,835
3. 2 + Expectations and parental involvement	1.14	.78	10,835	.71	.33	10,835
4. 3 if Course-taking variables not missing	1.36	.79	10,003	.86	.34	10,003
5. 4 + Climate, track, and course taking	-1.44	.75	10,003	.23	.36	10,003
6. 5 + Attended Catholic school in 8th grade	-1.98	1.03	10,003	-.34	.52	10,003
Reading Achievement Models						
1. Catholic school dummy variable only	3.98	.57	10,840	.91	.30	10,840
2. 1 + Family background and demographics	1.67	.52	10,840	.50	.32	10,840
3. 2 + Expectations and parental involvement	1.08	.51	10,840	.44	.32	10,840
4. 3 if Course-taking variables not missing	1.01	.50	10,003	.63	.30	10,003
5. 4 + Climate, track, and course taking	-.05	.55	10,003	.40	.34	10,003
6. 5 + Attended Catholic school in 8th grade	-.85	.78	10,003	.10	.48	10,003

Note: For all models, the data are weighted by the first follow-up questionnaire weight (F1QWT). The standard errors are robust standard errors, calculated with STATA's implementation of White's sandwich variance estimator modified to adjust further for clustering within schools.

both the mathematics test and the reading test.

The second panel of Table 1 presents six analogous models with the lagged variable specification adopted by Hoffer et al. (1985) in the second half of their article and as used in the more widely read book by Coleman and Hoffer (1987). Presented earlier as Equation 3b, the difference between the 12th-grade and 10th-grade test scores is regressed on the 10th-grade test score and the independent variables included in the analogous Models 1–6 reported in the first panel of Table 1. For this model specification, the Catholic school effect, as an estimate of the coefficient *d* in Equation 3b, declines from 1.31 to -.34 in mathematics as additional variables are added sequentially in Models 1–6. Likewise, the estimated effect on reading declines from .91 to .10. In comparison to the regression estimates presented in the first panel of Table 1, the estimated Catholic school effect in the second panel is smaller when few independent variables are included (Models 1–3) and larger when more independent variables are included (Models 5 and 6).

Taken together, the coefficient estimates in both panels of Table 1 show that one can obtain a wide range of estimates for the Catholic school effect on learning. As I discuss in the next section, without rigorous and explicit theoretical models of the process of learning and of how the Catholic school effect is generated, it is difficult to choose a preferred estimate from among the many estimates reported in Table 1.

How Definitive Are the Regression Estimates?

Regression techniques are the workhorse methodology of quantitatively oriented sociological research. Regression models, even the simple OLS techniques just used, are remarkably robust to biases induced by all manner of specification errors. As a result, regression models effectively pick up all the strong relationships that are present in available data.

The weaknesses of regression approaches emerge when effects are small and when regression specifications are tortured because the effects of interest are of paramount public importance. In research on school effects, both problems are present. School effects are indeed small, and they are certainly subject to intense public and political interest. For these reasons, sociological research on school effects could benefit from a less sanguine set of standards for interpreting regression estimates.

Why do regression estimates sometimes give the wrong answer? For a regression model to pick up a small effect, the model specification and its implicit functional form must be believable and trusted. The model presented in Equation 2 implicitly assumes that the amount of material learned by the senior year is some linear additive function of family background variables and an additive shock provided by the school. No theory of learning is so simple.

Classic behaviorist models of learning (see Bush and Estes 1959) specify complex alternative mechanisms for sequences of responses to learning trials. Mathematical models of learning were known to Coleman from his earliest involvement in educational research (see Coleman 1964:38). And although Coleman never specified a learning model that justified the prediction of test scores from a linear combination of family background characteristics and school shocks (see Sørensen and Morgan 2000), he did, in accordance with his early work on Markov chains and his proposals for longitudinal data analysis (Coleman 1964, 1981), provide an underlying model for the lagged achievement gain model in Equation 3b. Indeed, in Hoffer et al. (1985:89-91), he and his colleagues showed that (subject to restrictions on individual heterogeneity) the lagged test score model is a linearized reduced form model of two underlying rates (learning and forgetting) for the movement between two states (know and don't know) for each item on the cognitive test. Although plausible, Coleman's model is still limited by his desire to be able to estimate it with simple regression techniques (see Coleman 1981:8–9 for an explanation of his modus operandi in such situations).

Formal models of learning that can be tested with data are clearly needed (see Sørensen 1996; Sørensen and Hallinan 1977; Sørensen and Morgan 2000). For now, no such models are available that can be relied on to furnish a theoretically grounded functional form for a regression model that can be effectively deployed in school-effects research. As a result, regression techniques alone cannot be relied on to recover unbiased estimates of small effects, such as the Catholic school effect on achievement.

In the absence of a fine-grained model of the learning process, one would hope to be able to rely on a relatively simple and robust estimator. This hope, however, then leads to two even more fundamental questions: What is the simple quantity of interest that one should attempt to estimate? Can the typical survey data at our disposal provide a meaningful estimate of it?

Most researchers have centered attention on the estimation of a single Catholic school effect, implicitly defined as

(E1) The expected gain in achievement for a randomly selected student from the population if he or she was educated in a Catholic school instead of a public school.

Coleman and his colleagues focused attention on two more specific quantities of interest, implicitly defined as

(E2) The expected gain in achievement for a randomly selected public school student if he or she was instead educated in a Catholic school and

(E3) The expected gain in achievement that a randomly selected Catholic school student would forgo if he or she was instead educated in a public school.

When estimated with a single population regression model, such as those in Equations 1–3, estimates of E1, E2, and E3 are constrained to be equal. But when the effects of the independent variables X_1 through X_k and *Test*10 are allowed to vary across school sector (e.g., when separate regression models are estimated for each school sector), estimates of these three quantities will differ.[3] Unfortunately, if the separate regression equations are not justified by an explicit model of learning, there is no guarantee that sector-specific regressions will yield estimates that are any more easily interpretable than the estimates yielded by a single population model.

Even more ominous, perhaps, is that there may be many more than three Catholic school effects worthy of attention. Notice that in Equations 1–3, the Catholic school effect is parameterized by a constant *d* that is not subscripted by *i*. This implicit constraint can be relaxed in various ways to allow for the estimation of Catholic school effects for subgroups of the population and hence for an examination of the underlying heterogeneity of effects even within the two school sectors.

Random coefficient models are used in econometrics to allow for individual heterogeneity in regression coefficients. Accordingly, the coefficient *d* is subscripted by *i*:

$$(4) \quad Test12_i = a + d_i(Cath_i) + b_1X_{1i} + \ldots + b_kX_{ki} + e_i,$$

and a coefficient $\hat{\bar{d}}_i$ is estimated that is considered to be the mean effect of *Cath* on the test score (and suitable standard errors are produced that reflect variation in the coefficient *d* over *i*). Multilevel models are full generalizations of these random coefficient models that allow researchers to model variation in the effect *d* across individual-level characteristics, school-level characteristics, and interactions between the two. Taking this approach, the effective- schools branch of follow-up research on the Catholic school effect has used such models to clarify which types of private and public schools tend to generate the largest measured school effects and to probe for possible differential effects on learning within the two school sectors (see Bryk, Lee, and Holland 1993; Lee and Smith 1993, 1995). Thus, examinations of the heterogeneity of effects can be accommodated in a regression framework and elegantly in the multilevel models often used in school-effects research. In practice, however, the only heterogeneity that is explicitly discussed is that which can be modeled with the data that are available to serve as variables in X_1 through X_k. The counterfactual model I invoke in the next section starts with the assumption that heterogeneity of effects is pervasive and cannot necessarily be represented by a parsimonious parametric model estimated with the available data. The advantage of this approach is that important quantities of interest can be defined carefully before the available data are consulted. As a result, qualitative conclusions can be derived on the basis of reasonable theoretical statements even before any parameters are estimated.

CAUSAL EFFECT HETEROGENEITY

The Counterfactual Model of Causality

An extensive literature exists on the development within statistics and econometrics of notions of counterfactual causality.[4] In this presentation, I follow the structure and notation of Winship and Morgan (1999). I first introduce the main elements of the counterfactual model—the potential outcomes that are observable only on mutually exclusive subsets of a population. After I define an individual-level causal effect, I define the average causal effect and show why attempts to estimate it effectively with observational data often fail. I then emphasize the main value of the framework—its ability to motivate an analysis of patterns of causal effect heterogeneity that can clarify the meaning of standard regression estimates and suggest alternative semiparametric estimates of explicitly defined quantities of interest.

Potential Outcomes The counterfactual model presupposes that students have two theoretical scores on achievement tests—one that would be observed if they were educated in Catholic schools and one that would be observed if they were educated in public schools. (For now, assume that with respect to their effects on learning, all public schools are identical and all Catholic schools are identical.) Defining these potential outcomes respectively as Y_i^c and Y_i^p, the individual-level causal effect of Catholic schooling on achievement is then defined as

$$(5) \quad \delta_i = Y_i^c - Y_i^p.$$

Because only one of the potential outcomes can be realized and observed for each student, we cannot calculate the individual-level causal effect for any student.

If we have at our disposal a data set, $\{Y_i, C_i\}_{i=1}^n$, that is a simple random sample of size n from the population of high school students and where the variables Y_i and C_i are analogous, respectively, to $Test12_i$ and $Cath_i$ in Equations 1–4, then individual observations on the achievement test, Y_i, follow the simple observation rule:

$$(6) \quad Y_i = C_i Y_i^c + (1 - C_i) Y_i^p.$$

Thus, the distribution of the observed Y_i contains only half the information contained in the distributions of the theoretical potential outcome variables. And as a result, we cannot use the observed variables Y_i and C_i to identify the population distributions of either Y_i^c or Y_i^p.

Average Effects and the Standard Estimator Because we cannot calculate individual-level causal effects, we often focus attention on estimation of the average causal effect, defined as

$$(7) \; \bar{\delta} = \bar{Y}^c - \bar{Y}^p,$$

where $\bar{Y}^c$ and $\bar{Y}^p$ are population-level means of the corresponding individual-level potential outcomes. The average causal effect in Equation 7 is the most basic quantity of interest in studies of the Catholic school effect on achievement, defined earlier as E1, the expected learning gain that would be observed if a randomly selected student were educated in a Catholic school instead of a public school.

To understand how difficult it can be to estimate the average causal effect, consider the mechanism that generates the observable partition of students across the two school sectors. Students and their parents choose their schools, both within and between sectors, and their choices are subject to their preferences, financial constraints, and beliefs about the prospects of benefiting from Catholic schooling. Likewise, Catholic schools admit only some of the students from among their pools of applicants, after determining their own constraints and then estimating whom they expect will benefit most from attending Catholic schooling.

To represent these potential choice patterns more formally, assume that the population of students can be partitioned into two abstract sets of students C and P that can be substantively characterized as sets of students who are placed or who place themselves in alternative states defined as "attend Catholic school" and "attend public school." There are many potential partitions of the population across the two sets C and P, and the observational survey data that we typically analyze represent only one sample of students drawn from one possible population partition that the collective behavior of students, parents, and schools could generate.

For any partition across the sets C and P, two conditional population means exist: $\bar{Y}^c_{i\in C}$ and $\bar{Y}^p_{i\in P}$. Each of these conditional population means could, in theory, be calculated by a census of the entire population of students. The standard estimator of the average causal effect is formed by taking the difference between sample analog estimates of these conditional population means:

$$(8)\ \hat{\bar{\delta}} = \hat{\bar{Y}}^c_{i\in C} - \hat{\bar{Y}}^p_{i\in P}.$$

Thus, the standard estimate for a given data set is simply the observed difference between the sample means of the achievement test scores of Catholic school students and of public school students. In the regression framework presented earlier, an estimate of the parameter d in Equation 1 yields the same point-value estimate as $\hat{\bar{\delta}}$ in Equation 8.

When does the standard estimator in Equation 8 yield a consistent estimate of the true average causal effect in Equation 7? If we could randomly assign students to the two alternative school sectors, then we could justify the standard estimator as a consistent estimator of the true average causal effect. In essence, by manipulating the school-sector selection process, we would generate a partition in a set of students drawn from the population and then simply assert that for this partition, $\hat{\bar{Y}}^c_{i\in C} = \bar{Y}^c$ and $\hat{\bar{Y}}^p_{i\in P} = \bar{Y}^p$. As in most other school-effects research, such a randomization scheme is infeasible. Instead, students, parents, and schools generate the partition of the population over the sets C and P, and we can draw subjects only from a partitioned population to estimate the relevant conditional population means.

To understand why the standard estimator of the average causal effect may be poor when randomization is impossible, decompose the true average causal effect across those quantities that can be observed and those that cannot. While $\bar{Y}^c_{i\in C}$ and $\bar{Y}^p_{i\in P}$ are theoretically observable population means, their counterfactual analogs, $\bar{Y}^c_{i\in P}$ and $\bar{Y}^p_{i\in C}$, are inherently unobservable. These latter quantities are characterized as counterfactual because they exist in theory but cannot be verified through observation. The first counterfactual mean is the average outcome in the state "attend Catholic school" for those who, if sampled, would be observed to attend a public school. The second counterfactual mean is the average outcome in the state "attend public school" for those who, if sampled, would be observed to attend a Catholic school.

As in the case of randomization, the standard estimator yields a consistent estimate of the true average causal effect when $\bar{Y}^c_{i\in C} = \bar{Y}^c$ and when $\bar{Y}^p_{i\in P} = \bar{Y}^p$. By definition, these equalities will hold only if $\bar{Y}^c_{i\in C} = \bar{Y}^c_{i\in P}$ and $\bar{Y}^p_{i\in P} = \bar{Y}^p_{i\in C}$. Thus, with any set of data, randomized or observational, to justify the standard estimator as a consistent estimator of the true average causal effect, we must be able to assume that, on average, students who attend Catholic schools would have received the same scores on the test as those who attend public schools if those who attend Catholic schools had instead attended public schools (and vice versa).

Heterogeneity of the Causal Effect The causal effect of Catholic schooling may vary over individuals in the population, as suggested in the regression context by Equation 4. Most important, on average the causal effect of Catholic schooling may vary across the partition of those who choose (or who are chosen) to attend the alternative school sectors.[5]

To see why an explicit accounting of this heterogeneity is important, consider the following decomposition of the average causal effect defined earlier. First, let π equal the true proportion of the population that has chosen to attend a Catholic school. The average causal effect can then be decomposed into a weighted average of the average causal effect of attending a Catholic school for those who attend Catholic schools and the average causal effect of attending a Catholic school for those who attend public schools:

$$(9)\quad \bar{\delta} = \bar{Y}^{c} - \bar{Y}^{p}$$

$$= [\pi\, \bar{Y}^{c}_{i\in C} + (1-\pi)\, \bar{Y}^{c}_{i\in P}] - [\pi\, \bar{Y}^{p}_{i\in C} + (1-\pi)\, \bar{Y}^{p}_{i\in P}]$$

$$= \pi(\bar{Y}^{c}_{i\in C} - \bar{Y}^{p}_{i\in C}) + (1-\pi)(\bar{Y}^{c}_{i\in P} - \bar{Y}^{p}_{i\in P})$$

$$= \pi\bar{\delta}_{i\in C} + (1-\pi)\bar{\delta}_{i\in P}$$

Manipulating this decomposition, Winship and Morgan (1999) noted that there are two distinct sources of bias in the standard estimator of the average causal effect. For this application, they are the average difference between those who attend Catholic schools and those who attend public schools in (1) the baseline level of achievement that would exist if both groups of students attended public schools and (2) the average gain in achievement that would result if both groups of students attended Catholic schools instead of public schools. In particular,

$$(10)\quad \bar{Y}^{c}_{i\in C} - \bar{Y}^{p}_{i\in P} = \bar{\delta} + (\bar{Y}^{p}_{i\in C} - \bar{Y}^{p}_{i\in P}) + (1-\pi)(\bar{\delta}_{i\in C} - \bar{\delta}_{i\in P}).$$

The second source of bias is especially likely to be present if students self-select on the causal effect itself, as would be the case if the students who are more likely to benefit from attending a Catholic school recognize this potential gain and are disproportionately likely to enroll in a Catholic school.[6]

If there is self-selection on the causal effect itself, the best that we can hope to do is to estimate effectively the causal effect of Catholic schooling for those who typically choose to attend Catholic schools:

$$(11)\quad \bar{\delta}_{i\in C} = \bar{Y}^{c}_{i\in C} - \bar{Y}^{p}_{i\in C},$$

which was defined earlier as the effect E3 estimated by Coleman and his colleagues. If this is the only causal effect that we can effectively estimate, then we cannot effectively estimate either E1 or E2 because we cannot effectively estimate what the effect of Catholic schooling would be for those who typically attend public schools.

Stability of the Causal Effect Although I have referred to the Catholic school effect as a causal effect, much of the literature on counterfactual causality would refer to it as a treatment effect. An important assumption about the way in which a treatment effect is generated must be maintained to preserve the simplicity and power of the counterfactual framework. The stable unit treatment value assumption (SUTVA) requires that the potential outcomes of individuals would be unaffected by potential changes in the treatment statuses of other individuals. As a result, there can be no interference across treatments, and the treatment effect cannot depend on the number of individuals who are exposed to the treatment.

In this context, maintenance of SUTVA (also known as a no-macroeffect assumption) requires that the effectiveness of Catholic schooling not be a function of the number of students who enter the Catholic school sector. Is this assumption reasonable? For most school-effects researchers, I suspect not. For a variety of reasons—endogenous peer effects, capacity constraints, and so forth—most researchers would expect that the Catholic school effect would erode if a large number of public school students entered the Catholic school sector.

As a result, since there are good theoretical reasons to believe that macroeffects would emerge if Catholic school enrollments ballooned, it may be that we can estimate the causal effect of Catholic schooling only for those who would choose to attend Catholic schools, as in Equation 11, but also subject to the constraint that the proportion and composition of students who are educated in Catholic schools remains relatively constant. We may therefore be able only to estimate effectively:

$$(12)\quad \bar{\delta}^{*}_{i\in C} \equiv \bar{\delta}_{i\in C} \mid \text{SUTVA},$$

and this limitation is crucial for the policy relevance of estimates of the Catholic school effect on achievement, as is discussed in the Conclusions section.

Matching Estimates for Catholic School Students

The counterfactual model of causality, even though it highlights the severe limitations of observational survey data, suggests that for the Catholic school effect on achievement, we may be able to estimate successfully some important quantities of interest. In particular, even if school sector choices are based on students', parents', and schools' accurate perceptions of the potential benefits (or lack of benefits) of Catholic schooling, we may still be able to estimate the Catholic school effect for students who typically choose to attend Catholic schools (i.e., estimate $\bar{\delta}^*_{i \in C}$ from Equation 12). This is a theoretically important quantity, for if there is no Catholic school effect for Catholic school students, then most reasonable theoretical arguments would maintain that it is unlikely that there would be a Catholic school effect for students who typically attend public schools. And if policy interest was focused on whether Catholic schooling is beneficial for Catholic school students (and thus, for example, whether public support of transportation to Catholic schools is a benevolent government expenditure), then the Catholic school effect for Catholic school students is precisely the quantity we would want to estimate.

In this section, I use multivariate matching techniques to estimate $\bar{\delta}^*_{i \in C}$. A full introduction to matching is beyond the scope of this article. (For an accessible but nuanced introduction, see Smith 1997. For a full presentation, see Rubin and Thomas 1996, 2000 and Heckman, Ichimura, and Todd 1997, 1998).

I nonetheless offer a justification and skeletal outline of the specific procedure—matching with replacement on the estimated logit of the propensity score—that I used to generate the estimates presented later in Tables 4 and 5. The resulting estimates are not necessarily better than those offered by regression techniques. Their main value, from my perspective, is that they are more narrowly focused on specific quantities of interest. And they are more messy, in the sense that they reveal how much heterogeneity there really is in the data we typically analyze.

The Motivation for Multivariate Matching Techniques If we could observe the individual-level causal effect of Catholic schooling in Equation 5 for each Catholic school student, then estimation of $\bar{\delta}^*_{i \in C}$ would be trivial. We could simply take the mean across Catholic school students of these individual-level causal effects. Instead, with observational survey data, we can observe for each Catholic school student only a single value on the test score, $\hat{Y}_i$. By Equation 6, we know that for Catholic school students, the observed test scores on average correspond to values for Y_i^c. Multivariate matching techniques identify those public school students who are most similar to each Catholic school student, with respect to observable characteristics, in a global attempt to use the test scores of public school students to form estimates for each Catholic school student's unobservable Y_i^p.

Matching estimators are feasible and most easily defendable in applications in which treatment selection is nondeterministic. For Catholic school attendance, such an assumption is reasonable, and I therefore maintain that decisions on the selection of school sector are fundamentally stochastic. In other words, conditional upon all observable predictors of the selection of school sector, the decision to attend a Catholic school is a random draw (e.g., from a Bernoulli distribution). A direct implication of this assumption is that each observed student, even though observed attending school in only one school sector, has complementary nonzero predictive probabilities of attending both a Catholic school and a public school. The goal of the form of matching that I used is to match to each Catholic school student one or more public school students with the same predictive probability of attending a Catholic school.

To launch a matching routine, one must first assemble a set of variables, S_1 through S_m, that predict school sector selection. These variables are not thought of as determinants of learning, although they may be those as well. With NELS data, only some of the variables utilized in the regression models, such as family background and basic demographic characteristics, naturally belong among the variables S_1 through S_m. I would argue that other variables, such as educational expectations and academic climate measured in the 10th grade, are not reasonable candidate variables for S_1 through S_m because they are influenced to an unknown degree by Catholic school attendance itself (see Lieberson 1985; Rosenbaum 1984a). This position is, of course, debatable.

For notational convenience, situate these "pretreatment" variables S_1 through S_m in individual-specific row vectors, $\mathbf{S}_i$. If, as a set of predictor variables, $\mathbf{S}_i$ is sufficiently exhaustive of determinants (other than anticipation of the potential causal effect itself) of sector selection, then it may be permissible to condition on values of $\mathbf{S}_i$ and declare that

$$\text{(A1)}\quad \bar{Y}^{p}_{i\in C} \mid \mathbf{S}_i = (\bar{Y}_i \mid C_i = 0, \mathbf{S}_i).$$

Assumption A1 asserts that for every subpopulation of students with a distinct set of values for the variables S_1 through S_m, the observable mean test score ($\bar{Y}_i$) of public school students (those for whom it would be observed that $C_i = 0$) is equal to the unobservable counterfactual mean of test scores ($\bar{Y}^{p}_{i\in C}$) for Catholic school students if they had instead attended public schools.[7]

If Assumption A1 is valid, then we can define a consistent estimator of $\bar{\delta}^{*}_{i\in C}$ which is conditioned on distinct combinations of values on the variables in $\mathbf{S}_i$:

$$\begin{aligned}\text{(M1)}\ \hat{\bar{\delta}}^{*}_{i\in C} \mid \mathbf{S}_i &= (\bar{Y}^{C}_{i\in C} - \bar{Y}^{p}_{i\in C}) \mid \mathbf{S}_i \\ &= (\bar{Y}^{C}_{i\in C} \mid \mathbf{S}_i) - (\bar{Y}^{p}_{i\in C} \mid \mathbf{S}_i) \\ &= (\bar{Y}_i \mid C_i = 1, \mathbf{S}_i) - (\bar{Y}_i \mid C_i = 0, \mathbf{S}_i).\end{aligned}$$

We can then average over estimates within these strata to form a consistent estimate of $\bar{\delta}^{*}_{i\in C}$ when necessary. The goal of multivariate matching is thus to stratify the sample along a set of predictors of school-sector selection that can justify Estimator M1 as a consistent estimator of a stratified version of $\bar{\delta}^{*}_{i\in C}$.

Matching on the Propensity Score The practical problem that emerges with this approach is known as the curse of dimensionality. If we need many variables in $\mathbf{S}_i$ to characterize sector selection and if we have a finite sample (as we always do), it is unlikely that we will be able to find at least one public school student who has the exactly equivalent row vector $\mathbf{S}_i$ of each Catholic school student.[8]

One potential solution is to match on a single dimension that has a claim to be an optimally weighted function of the variables in $\mathbf{S}_i$. Following Rosenbaum and Rubin (1983), the favored single dimension in the matching literature is known as the propensity score.[9] In this context, it is defined as the predictive probability of attending a Catholic school, given the variables that predict sector selection:

$$\text{(13)}\quad P(\mathbf{S}_i) \equiv Pr(i \in C \mid \mathbf{S}_i).$$

The definition in Equation 13 states that the propensity score $P(\mathbf{S}_i)$ is the true predictive probability (between 0 and 1) that a student with characteristics $\mathbf{S}_i$ would choose or would be chosen to enroll in a Catholic school ($i \in C$). Collectively, the propensity score values are a set of true predictive probability values that do not necessarily trace any familiar probability density function.

The logic of the propensity-score approach—as a potentially feasible simplification of the more general multivariate matching approach justified by Assumption A1 and executed in Estimator M1—is that Catholic school students and public school students with the same propensity score can be treated as if they are equivalent in all other respects relevant for estimating the Catholic school effect for Catholic school students.[10] Rosenbaum and Rubin (1983) showed that the goal of all matching techniques is to balance the distributions of all relevant pretreatment variables $\mathbf{S}_i$ across the two groups under comparison. They showed that one can match on the propensity score, $P(\mathbf{S}_i)$ instead of $\mathbf{S}_i$ itself, so that in this context while maintaining Assumption A1, one can execute the following estimator:

$$\begin{aligned}\text{(M2)}\ \hat{\bar{\delta}}^{*}_{i\in C} \mid P(\mathbf{S}_i) &= (\hat{\bar{Y}}^{C}_{i\in C} - \hat{\bar{Y}}^{p}_{i\in C}) \mid P(\mathbf{S}_i) \\ &= (\hat{\bar{Y}}^{C}_{i\in C} \mid P(\mathbf{S}_i)) - (\hat{\bar{Y}}^{p}_{i\in C} \mid P(\mathbf{S}_i)) \\ &= (\hat{\bar{Y}}_i \mid C_i = 1, P(\mathbf{S}_i)) - \\ &\quad\ (\hat{\bar{Y}}_i \mid C_i = 0, P(\mathbf{S}_i)).\end{aligned}$$

They then offered a proof that matching on the propensity score, as in Estimator M2, rather than simultaneously on all variables in $\mathbf{S}_i$, as in Estimator M1, is all that is necessary to provide balance on the dimensions of $\mathbf{S}_i$.[11]

To implement a perfect propensity-score matching estimator, one must have access to the propensity score, $P(\mathbf{S}_i)$. Unfortunately, the propensity score is, by definition, an unknown true quantity that must be estimated. Although any parametric or nonparametric estimator of the propensity score is possible, a generic logit model is usually chosen:

$$(14)\quad \Pr[C_i = 1|\mathbf{S}_i] = \frac{\exp(\mathbf{S}_i\phi)}{1 + \exp(\mathbf{S}_i\phi)}.$$

In the following analysis, I adopt this convention. Moreover, I also follow convention and match on the index of the logit model (i.e., the logit of the predicted probability), $e(\mathbf{S}_i) = \mathbf{S}_i\hat{\phi}$, which is a monotonic transformation of the propensity score that avoids the compression of the probability scale near 0 and 1. Although this approach to estimating the propensity score is consistent with tradition, there are few guidelines to assess how much bias may result from reliance on the logit functional form (see Rubin and Thomas 2000). The specific assumption that I invoke to estimate $\bar{\delta}^*_{i\in C}$ is therefore that

$$(A2)\quad (\bar{Y}^p_{i\in C}|e(\mathbf{S}_i)) = (\bar{Y}_i\,|C_i = 0, e(\mathbf{S}_i)),$$

which differs from Assumption A1 in that $e(\mathbf{S}_i)$ is both an estimated quantity and a single dimension.

With this approach, there are several reasons for caution. First, because $e(\mathbf{S}_i)$ is an estimated quantity, Assumption A2 can be true in expectation only over repeated samples. Thus, one could replace $e(\mathbf{S}_i)$ in Assumption A2 with its expectation over repeated sample estimates, $E(e(\mathbf{S}_i))$, but such manipulation is a formalist evasion of a potentially serious practical problem. Instead, I assess the sensitivity of the matching estimators to this possible single-sample threat using pooled within-strata regression adjustments, just as an experimentalist would adjust for chance differences in the distributions of important covariates across randomized treatment and control groups.[12]

Second, although propensity-score methods avoid the need for an explicit model of learning, the true functional dependence of learning on $\mathbf{S}_i$ is not irrelevant. Although not widely appreciated in the 1980s, when the propensity-score literature was first developing, it is now more generally recognized that propensity-score matching tends only to balance the means of the variables in $\mathbf{S}_i$. This limitation can be a problem. For example, if the propensity score is a function of S_i (or, in the case here, the logit of the propensity score is a linear function of S_i), but learning is a quadratic function of S_i, then propensity-score methods tend to balance cases on the mean of S_i but not on the mean of the square of S_i.

Third, Assumption A2 may not be valid because the variables in $\mathbf{S}_i$ may not characterize sector selection completely enough to justify the equality even in the less restrictive Assumption A1. With the data analyzed later, this is a distinct possibility, especially since I include in $\mathbf{S}_i$ only variables (or functions of variables) also used by Hoffer et al. (1985). For this article, I restrict $\mathbf{S}_i$ in this way for three reasons: (1) to enable a direct assessment of the contribution of an alternative technique, (2) to show how matching techniques can be used alongside regression techniques to assess whether there is interpretable causal effect heterogeneity beneath a regression estimate, and (3) because there simply are few other plausible variables within the NELS data set that could reliably serve as sector-selection predictors.

As I discuss later, matching estimates that utilize more extensive sets of sector-selection predictors will be needed to resolve the patterns documented in the next section. My sense is that great leaps forward will be possible only when more informative data are available. Given these qualifications and limitations, in many ways no less stringent than those that afflict regression models, the effort to generate propensity-score matching estimates seems worthwhile to place the regression estimates within a more broad set of possible estimates.

Results of Propensity-Score Matching To obtain the logit of the estimated predictive probability of attending a Catholic school, I first estimated logit models separately for the mathematics and reading tests (because slightly different samples of students have nonmissing values on both the 10th- and 12th-grade tests). The models contained 62 parameters: a constant, main effects for the family background and basic demographic variables listed in Table A1, the first follow-up sample weight (F1QWT), and two-way interactions between each of the race dummy variables with all other variables, socioeconomic status (SES) with all other variables, and the weight with all other variables. For a second set of matching estimates—parallel to the strategy in regression analysis of moving from the specification in Model 2 to the more exhaustive specification in Model 3—I then reestimated these logit models to obtain even more precisely estimated propensity scores, specifying the respective 10th-grade test scores as a variable in $\mathbf{S}_i$.

In line with the guidance of Rubin and Thomas (1996), statistical significance of the logit coefficients was not an important criterion in settling on the specification of higher-order interaction terms (or in deciding not to use higher-order polynomial terms). Instead, achieving the best possible balance on the means and standard deviations of the family background and demographic characteristics was the main criterion for selecting the specification. Moreover, because no NELS Catholic schools were designated as rural, I effectively performed a perfect match on nonrural status by dropping all public school students who attended rural schools before I estimated the logit models.

Table 2 presents selected quantiles of the distributions of $e(\mathbf{S}_i)$, separately for public school students and Catholic school students (and for specific samples that correspond to the four sets of models presented later in the four panels of Tables 4 and 5). Although there is considerable overlap in the range of $e(\mathbf{S}_i)$ across public school students and Catholic school students, the distribution for Catholic school students is, of course, shifted to the right of that of public school students. For example, the Catholic school student at the median value of $e(\mathbf{S}_i)$ for Catholic school students has a value for $e(\mathbf{S}_i)$ that is greater than the corresponding value of $e(\mathbf{S}_i)$ for the public school student at the 80th percentile of the public school students' distribution. Note that there are some public school students who have values for $e(\mathbf{S}_i)$ that are smaller than the minimum value estimated for all Catholic school students. Because I attempted to estimate the Catholic school effect for Catholic school students only, the existence of this group of public school students presented no problems. However, also notice that a few Catholic school students have values for $e(\mathbf{S}_i)$ that are greater than the maximum value estimated for all public school students. As I show later, the presence of these students complicates estimation of the Catholic school effect for all Catholic school students, for there are simply no public school students who are similarly prone to Catholic school attendance as are these students.

With $e(\mathbf{S}_i)$ in hand, a matching algorithm must be invoked to take each Catholic school student and match to that student all public school students who are similar with respect to $e(\mathbf{S}_i)$. Because more than one public school student may be matched to each Catholic school student, this type of multiple matching is sometimes referred to as stratification on the propensity score, with each Catholic school student and his or her matched public school students forming unique strata. I used two related matching algorithms, both of which are summarized in Table 3, alongside descriptions of two types of estimation procedures used for Tables 4 and 5.

The complete match described in the first panel of Table 3 ensures that all Catholic school students are matched to at least one public school student, even though some of the matches may be poor in the sense that for each additional forced singleton match (Step 2) the distance between $e(\mathbf{S}_i)$ for the Catholic school student and his or her matched public school student may not be as small as is desired.

In contrast, the common-support match ensures that only close matches are utilized. But with this alternative matching algorithm, Catholic school students who do not have at least one close match are dropped from the resulting stratified data set, and inference for estimates produced from analysis of the stratified data set must be restricted to Catholic

Table 2. Distributions of the Estimated Logit of the Propensity Score for Attending Catholic School

Specification of Propensity Score	*N*	Minimum	1st Percentile	20th Percentile	40th Percentile	Median	60th Percentile	80th Percentile	99th Percentile	Maximum
For Mathematics Achievement										
Family background and demographics										
Public school students	6,197	-41.103	-8.561	-4.519	-3.517	-3.168	-2.875	-2.033	-.062	1.780
Catholic school students	719	-9.083	-5.357	-2.264	-1.234	-.832	-.575	-.195	.601	2.870
Family background and demographics and 10th-grade test score										
Public school students	6,197	-41.089	-8.660	-4.560	-3.564	-3.196	-2.895	-2.057	.018	1.611
Catholic School Students	719	-8.887	-5.539	-2.261	-1.198	-.853	-.578	-.186	.776	2.673
For Reading Achievement										
Family background and demographics										
Public school students	6,210	-41.269	-8.529	-4.536	-3.527	-3.172	-2.880	-2.042	-.069	1.590
Catholic school students	716	-8.926	-5.331	-2.290	-1.237	-.833	-.567	-.200	.611	2.913
Family background and demographics and 10th-grade test score										
Public school students	6,210	-40.412	-8.598	-4.548	-3.570	-3.202	-2.892	-2.044	-.014	1.620
Catholic school students	716	-8.740	-5.368	-2.286	-1.234	-.882	-.583	-.179	.732	3.126

Note: Because no Catholic schools are designated as rural schools, all public school students who attend rural schools were dropped from the analysis before the logit models were estimated.

Table 3. Alternative Matching Algorithms and Effect Estimators

	Complete Match	Common-Support Match
Matching Algorithm		
Step 1: Caliper match	For each Catholic school student with $e(\mathbf{S}_{i \in C}) = k_i$, identify all public school students for whom $e(\mathbf{S}_{i \in P}) = \in [k_i - .01, k_i + .01]$. Post each Catholic school student and his or her matched public school students to the new data set with a unique stratum identification number. Because some public school students will be matched to more than one Catholic school student (i.e., assigned to more than one strata), the new dataset will have duplicated public school students and will therefore have an artificially inflated *N*.	Same.
Step 2 : Additional forced-singleton match	For each Catholic school student with $e(\mathbf{S}_{i \in C}) = k_i$ and without any matched public school students from Step 1, select the single public school student whose value for $e(\mathbf{S}_i)$ is closest to k_i, and post each of these Catholic school students and his or her single matched public school student to the new data set with a unique stratum identification number.	Skip. (Catholic school students whose values of $e(\mathbf{S}_i)$are not similar enough to any public school students (i.e., whose caliper does not contain any public school students) are not posted to the new data set and are therefore dropped from further analysis).
Effect Estimation		
Unadjusted	Estimate a fixed effect model with a fixed effect for each stratum (i.e., an OLS regression model with separate dummy variables for all but one of the strata, as in STATA's areg command). Adjust standard errors for true sample size: number of Catholic school students plus the number of public school students who are ever matched.	Same.
Regression adjusted	Estimate a fixed effect model with a fixed effect for each stratum and additional independent variables, such as family background and demographic characteristics and a lagged test score. Adjust standard errors for true sample size: number of Catholic school students plus the number of public school students who are ever matched.	Same.

Table 4. Propensity-Score Matching Estimates of the Catholic School Effect for Catholic School Students

	Complete Match (Caliper match with additional forced singleton matching)						Common Support Match (Caliper match)					
	Unadjusted		Regression Adjusted Within Strata				Unadjusted		Regression Adjusted Within Strata			
	Catholic School Effect	*SE*	Catholic School Effect	*SE*	True *N*	Inflated *N*	Catholic School Effect	*SE*	Catholic School Effect	*SE*	True *N*	Inflated *N*
Mathematics												
All strata	3.650	.917	3.345	.817	4,430	11,169	3.116	.941	3.019	.836	4,398	11,117
1st quintile	3.207	1.567	2.710	1.367	2,429	4,709	3.025	1.612	2.379	1.403	2,425	4,705
2nd quintile	5.977	2.026	5.698	1.830	833	2,071	5.977	2.026	5.698	1.830	833	2,071
3rd quintile	3.593	2.629	3.380	2.553	525	1,750	3.593	2.629	3.380	2.553	525	1,750
4th quintile	.394	3.087	.944	2.884	375	1,835	.394	1.139	.944	2.884	375	1,835
5th quintile	6.130	3.090	4.319	2.805	268	804	2.589	1.154	3.182	1.241	240	756
1st-3rd quintiles	4.078	1.062	3.723	.942	3,787	8,530	4.033	1.073	3.632	.950	3,783	8,526
4th-5th quintiles	2.825	2.267	2.779	2.096	643	2,639	1.151	.918	1.737	2.193	615	2,591
Reading												
All strata	2.208	.630	1.911	.580	4,514	11,225	1.647	.642	1.527	.589	4,489	11,183
1st quintile	2.234	1.067	1.748	.991	2,508	4,643	2.234	1.067	1.748	.991	2,508	4,643
2nd quintile	2.100	1.495	2.023	1.363	838	2,135	2.100	1.495	2.023	1.363	838	2,135
3rd quintile	1.499	1.781	2.661	1.689	544	1,820	1.499	1.781	2.661	1.689	544	1,820
4th quintile	2.064	2.172	2.695	2.057	360	1,806	2.064	2.172	2.695	2.056	360	1,806
5th quintile	3.629	2.180	3.019	2.194	264	821	-.713	1.028	.381	2.260	239	779
1st-3rd quintiles	1.921	.733	1.626	.672	3,890	8,598	1.921	.733	1.626	.672	3,890	8,598
4th-5th quintiles	2.756	1.573	2.898	1.486	624	2,627	1.051	1.658	1.674	1.549	599	2,585

Note: Because public school students who attend rural schools were dropped from the analysis sample before the propensity score was estimated, the students are perfectly matched on rural-nonrural, although not necessarily perfectly matched on urban-suburban. Standard errors are only approximate because they are not adjusted for the clustering of students within schools. The strata are weighted by the weight (F1QWT) attached to each Catholic school student in that strata.

Table 5. Propensity-Score Matching Estimates of the Catholic School Effect for Catholic School Students (with the 10th-Grade Test Score Specified as an Underlying Dimension of the Propensity Score)

	Complete Match (Caliper match with additional forced singleton matching)						Common Support Match (Caliper match)					
	Unadjusted		Regression Adjusted Within Strata				Unadjusted		Regression Adjusted Within Strata			
	Catholic School Effect	*SE*	Catholic School Effect	*SE*	True *N*	Inflated *N*	Catholic School Effect	*SE*	Catholic School Effect	*SE*	True *N*	Inflated *N*
Mathematics												
All strata	1.882	.873	1.152	.361	4,374	10,829	1.826	.888	.971	.365	4,353	10,791
1st quintile	3.259	1.545	.900	.634	2,411	4,646	3.259	1.545	.900	.634	2,411	4,646
2nd quintile	4.011	1.925	1.565	.774	851	2,059	4.011	1.925	1.565	.774	851	2,059
3rd quintile	.355	2.701	.639	1.126	480	1,728	.355	2.701	.639	1.126	480	1,728
4th quintile	1.131	2.812	.513	1.276	357	1,605	1.131	2.812	.513	1.276	357	1,605
5th quintile	.628	2.172	2.298	1.217	275	791	-.159	2.406	1.322	1.302	254	753
1st-3rd quintiles	2.401	1.045	1.103	.422	3,742	8,433	2.401	1.045	1.103	.422	3,742	8,433
4th-5th quintiles	.900	1.873	1.203	.893	632	2,396	.617	1.968	.681	.921	611	2,358
Reading												
All strata	1.218	.616	.782	.367	4,487	10,852	.970	.632	.626	.375	4,457	10,804
1st quintile	1.320	1.077	.523	.623	2,487	4,629	1.490	1.092	.685	.630	2,485	4,627
2nd quintile	.180	1.391	-.151	.810	853	2,144	.180	1.391	-.151	.810	853	2,144
3rd quintile	2.219	1.871	2.394	1.089	523	1,750	2.219	1.871	2.394	1.089	523	1,750
4th quintile	.968	1.938	.449	1.343	363	1,627	.968	1.938	.449	1.343	363	1,627
5th quintile	.948	1.540	1.512	1.302	261	702	-1.911	1.730	.225	1.373	233	656
1st-3rd quintiles	1.353	.737	.935	.425	3,863	8,523	1.415	.741	.988	.426	3,861	8,521
4th-5th quintiles	.959	1.330	.844	.958	624	2,329	-.004	1.416	.095	1.000	596	2,283

Note: See the note to Table 4.

school students who have comparable counterparts among public school students.

Given the variables used in $\mathbf{S}_i$ to estimate the logit of the propensity score, it is most natural to compare the four sets of matching estimates presented in Tables 4 and 5 to the regression coefficient estimates from Model 2 in Table 1. Even though I make such comparisons, it should be noted that the propensity-score-matching estimates stand on their own as viable estimates. The matching estimates are not, therefore, corrected regression estimates; instead, they are alternative estimates that are themselves subject to their own set of weaknesses.

Consider first the complete-match estimates for all strata. For these estimates of $\bar{\delta}^*_{i\in C}$, the unadjusted estimates are larger—20–50 percent larger—than the corresponding regression estimates. When these estimates are regression adjusted through a pooled within-strata estimator, the estimates decline but remain larger than the regression estimates reported in Table 1.

Now consider the common-support-match estimates for all strata. For these estimates (as reflected in the column for the number of respondents), Catholic school students without good matches are dropped from the analysis, and most but not all these students are in the top quintile of the distribution of $e(\mathbf{S}_i)$. These omitted students tend to score very well on the achievement tests, for by all standards they are from advantaged social origins, and this may be why the common-support-match estimates are smaller than the complete-match estimates.

Which set of matching estimates should we consider more worthy? Assuming that the variables in $\mathbf{S}_i$ are exhaustive enough to justify Assumption A1, the complete-match estimates should be regarded as upper-bound estimates of $\bar{\delta}^*_{i\in C}$, and the common-support-match estimates should be regarded as lower-bound estimates of $\bar{\delta}^*_{i\in C}$. With this interpretation, the common support-match estimates are at least as large as the simple regression estimates (and usually larger), and the complete-match estimates are considerably larger than the regression estimates. With a larger sample of data, I would therefore expect that a more complete propensity-score matching scheme (i.e., one for which the common-support and complete-match samples are exactly equivalent) would yield an estimate somewhere between the complete match and common-support match estimates reported in Tables 4 and 5.[13]

Perhaps even more interesting are the estimates for quintiles of the distribution of strata, based on the ranking of Catholic school students by $e(\mathbf{S}_i)$. Standard regression estimators do not reveal such quantities and imply that without further parameterization, there would be no interpretable pattern to the variation in their estimates if the analytic sample were subdivided in any way.

In contrast, for the propensity-score matching estimators in Tables 4 and 5, there is considerable variation in estimates of the average causal effect for Catholic school students with different propensities for attending Catholic schools. Beneath the noise that represents sampling error, it appears as if the Catholic students who are least likely to be enrolled in Catholic schools (as predicted by the variables in $\mathbf{S}_i$), are the most likely to benefit from having attended a Catholic school. This pattern is more convincing when (1) it is recognized that the high level of the average effect estimate for the fifth quintile in the complete match is likely biased by the presence of relatively poorly matched public school students and (2) sampling noise is diminished by collapsing the first through third quintiles and the fourth and fifth quintiles into only two groups, as in the bottom two lines of each panel in Tables 4 and 5.

How do these Catholic school students differ in their underlying distributions across the variables in $\mathbf{S}_i$? Table 6 presents means for these variables and for the mathematics achievement test for two groups of Catholic school students: those below the 60th percentile of the distribution of the propensity score for Catholic school students and those above the 60th percentile.

Several patterns are clear. Although students who are the least likely to attend Catholic schools are the most likely to benefit from having done so, on an absolute scale of achievement, the students who are more likely to attend Catholic schools tend to score higher on the mathematics test in both their sophomore

Table 6. Means for Catholic School Students by Propensity to Attend a Catholic School (for a Selected Model of Mathematics Achievement)

	1st-3rd Quintiles (N = 431)	4th and 5th Quintiles (N = 288)
Achievement Test Scores		
1992 Mathematics (IRT scaled)	52.783	57.910
1990 Mathematics (IRT scaled)	47.243	52.531
Family Background and Other Demographic Characteristics		
Socioeconomic status	.253	.550
Urban	.721	.999
Suburban	.279	.001
Northeast	.264	.364
North Central	.226	.458
South	.319	.049
West	.191	.129
Number of siblings	2.132	2.278
Have own bedroom	.828	.730
Two-parent family	.786	.950
White	.704	.811
Asian	.038	.102
Hispanic	.150	.043
Black	.100	.044
American Indian	.008	.000
Learning or physical disability	.074	.056

Note: Relevant sample is the Complete Match sample for the mathematics model with the 10th-grade test score as a predictor in the propensity score (i.e., those in the fourth line of Table 4 and in the Mathematics, Complete Match panel in Table 7). Data are weighted by the first follow-up questionnaire weight (F1QWT).

and senior years. Considering the means of the other variables in Table 6, this baseline achievement difference is not surprising. Students from the fourth and fifth quintiles of the propensity-score distribution are more likely to have been raised in families with two parents and with high levels of SES. More members of this group of students are also white and Asian. Nonetheless, this pattern means that those students who most benefit from Catholic schooling tend to be low-SES students who are disproportionately likely to identify themselves as black or Hispanic.

With the imposition of a patterned interpretation on the heterogeneity across quintiles in the matching estimates presented in Tables 4 and 5, a dose of honesty is in order. The findings are perhaps still too cloudy to justify the clean accounting that I have given them. In particular, there is not as much consistency across mathematics and reading achievement as one might hope in order to have more confidence that the modeled propensity score is on target. And, even more deeply, it is simply hard to know whether departures from the basic story can be explained away as sampling error. If either of these two sources of potential noise is found in future research to be an important part of the heterogeneity of the estimates I offer here, then the explanatory narratives I offer in the Conclusions will be less convincing, including the argument of other researchers that Catholic schools now more commonly approach the model of a common school than do public schools.

CONCLUSIONS

By focusing attention on the potential pervasiveness of causal effect heterogeneity, the counterfactual model of causality clearly demonstrates some of the limitations of observational data. More optimistically, however, the counterfactual model also demonstrates that some important quantities of interest can be effectively estimated with observational survey data.

Under the reasonable theoretical assumption that students, parents, and schools make school enrollment decisions partly on the basis of the prospects for each student's learning, then it may be impossible to use observational data to provide a consistent estimate of what the Catholic school effect would be for students who typically attend public schools. Nonetheless, this impossibility does not mean that one cannot estimate what the Catholic school effect is for those who typically choose to attend Catholic schools or that this quantity is not worth estimating.

With a propensity-score-matching estimator, I provided estimates of this "treatment effect for the treated" parameter, which suggest that the Catholic school effect for Catholic school students is larger than is suggested by the standard constant-coefficient, single-population regression model offered in the first portion of the article. Moreover, by using the propensity-score-matching estimator, I was able to examine whether patterns of causal effect heterogeneity have any interpretable relationship with the propensity to attend Catholic school. Indeed, it appears more likely than not that there is interpretable variation, suggesting that students who are the least likely to attend Catholic schools on the basis of their observed characteristics are the most likely to benefit from having done so.

This evidence of patterned heterogeneity of the average causal effect is supportive of at least one of four different underlying narratives:

1. For the *common-school narrative*, the effect of Catholic schools may indeed vary, and it may vary in ways that are typically considered to be virtuous. In particular, students from disadvantaged social backgrounds—black and Hispanic students from low-SES families— are most likely to have benefited from having attended Catholic schools. Thus, Catholic schools may be common schools that distribute opportunities for learning more equitably than do public schools, as stressed by Coleman and his colleagues and more completely by Bryk et al. (1993).
2. For the *differential-sacrifice narrative*, disadvantaged Catholic school students may resolve to work harder than their relatively advantaged peers simply because they are more motivated by knowing that they and their parents are making a genuine sacrifice.
3. The *better-alternatives narrative* stresses a demographic reality, not the apparent equity of achievement that is produced by Catholic schools or a behavioral response to differential sacrifice. Catholic schooling is particularly beneficial to students who have poor public schooling alternatives, especially students from families who are not able to afford to live in public school districts with the best public schools (see Neal 1997).
4. For the *binding-constraint narrative*, it may be that there is no genuine heterogeneity whatsoever or, at least, no heterogeneity in a true effect that is related to the propensity to attend a Catholic school. Instead, it may be that there is differential responsiveness to selection on an accurate perception of a student's likely benefit from Catholic schooling (e.g., some direct function of Y_i^c - Y_i^p). For some low-SES families for whom tuition at a Catholic school represents a genuine financial sacrifice, the only students who enroll in Catholic schooling may be those who are especially likely to benefit from enrolling. In contrast, among high-SES families for whom tuition is not a substantial financial sacrifice, even students who are not likely to benefit from attending Catholic schooling may enroll in a Catholic school.

Sorting out the relative plausibility of these four narratives should be the next goal of research on the Catholic school effect on achievement.

As for the public policy relevance of the estimates, it should first be stipulated that the

Catholic school effect on achievement is sociologically interesting and worthy of careful study, regardless of its apparent policy relevance. Learning how and why different types of schools generate different outcomes is intellectually satisfying, as has been demonstrated in many of Coleman's novel essays on functional and value communities and in his proposal that intergenerational closure should be seen as a form of social capital. But given all the complications inherent in the analysis of observational data, do estimates of the Catholic school effect from survey data have great policy relevance, as proponents of school choice and voucher programs often claim?

Completely apart from empirical analysis, the counterfactual model reveals important policy implications. With available survey data but without a trusted model of learning, there is almost no way to provide a consistent estimate of how well public school students would perform on standardized tests if they were instead enrolled in existing Catholic schools. If there is self-selection to an unknown degree on the causal effect itself (and it seems almost undeniable that there should be at least some self-selection), then even if we can identify strata of observationally equivalent public school students and Catholic school students, we will be able to estimate effectively only the Catholic school effect for the Catholic school students.

Even if we could generate consistent estimates of the Catholic school effect for public school students if they had been educated in the current Catholic school sector, the available survey data cannot tell us how Catholic schools would respond to an influx of former public school students, armed with vouchers or not. Thus, we would (or at least should) have little confidence in our ability to use any such seemingly perfect estimates to infer whether or not the Catholic school effect on achievement would persist for a new distribution of students across school sectors, as would be the outcome of any substantial policy intervention.

From a relatively objective policy perspective, the main implication of all past research on the Catholic school effect should be this: While there is evidence that there is a Catholic school effect on achievement for students who chose or who were chosen to attend Catholic schools in the past, this evidence has little bearing on policy debates about how students who are currently educated in public schools should be educated in the future.

APPENDIX A

Details of the Data and of the Regression Modeling Strategy

The goal of the regression models partially reported in Table 1 is to obtain a set of estimates that matches those that I suppose Hoffer et al. would have obtained if they had analyzed the NELS data with the same basic procedures adopted for their 1985 article. In this appendix, I present the particulars of the regression models I offer and some basic descriptive comparisons of differences in the test scores recorded for HS&B in the early 1980s and for NELS in the early 1990s. In the final section of the appendix, I assess the consequences of alternative procedures for handling missing data—in Hoffer et al. (1985), in the research of those who challenged Coleman and his colleagues, and in the regression models that I offer in this article.

Comparability of Data Like its predecessor HS&B, NELS (National Center for Education Statistics 1996) is a two-stage stratified random sample of students nested within schools. The NELS, however, is based on a substantially more complicated longitudinal design. As a result, there may be important differences between the HS&B and NELS samples that should prevent the reader from considering the analysis reported here as any indication of whether Hoffer et al. (1985) obtained reasonable estimates from their analysis of HS&B. Moreover, since there may also have been substantial changes in enrollment patterns in Catholic schools between 1980 and 1990 (see Bryk et al. 1993), any attempt to use NELS to evaluate the specific parameter estimates of Hoffer et al. would be misdirected.

Analysis Sample, Variables, and Basic Descriptive Statistics I initially selected NELS respondents who were enrolled as sophomores in 1990 in either a public or a Catholic school and who were also enrolled as seniors in a public or a Catholic school in 1992. I therefore excluded from the analysis all NELS students who were enrolled in other types of private schools, sophomores who were not enrolled as seniors in 1992 (either because they graduated early or dropped out of school between 1990 and 1992), and seniors who were added to the sample as freshened students in 1992. From this theoretical analysis sample, I dropped an additional 118 respondents who transferred between the public school and Catholic school sectors between their sophomore and senior years.[14]

I selected variables from the NELS data set that match those used by Hoffer et al. (1985) as closely as possible. Because of the similarities of the questionnaires in the two surveys, both of which were sponsored by the National Center for Education Statistics, this selection was feasible.[15] The means and standard deviations of the variables used are presented in Table A1, along with the NELS source variable and item-specific valid numbers of cases.

Table A2 presents mean test scores in mathematics and reading for Catholic school students and public school students. The first two rows are drawn from Hoffer et al. and present test score means of HS&B sophomores in 1980, test score means of HS&B seniors in 1982, and mean achievement gains between the sophomore and senior years. For a baseline comparison with Hoffer et al.'s findings, the third and fourth rows present analogous test score means and mean achievement gains for NELS sophomores in 1990 and NELS seniors in 1992.

For both HS&B and NELS, the means of the test scores are higher for students who attend Catholic schools, in both the sophomore and senior years. Furthermore, the mean achievement gain between the sophomore and senior years is higher for those who attend Catholic schools.

The seventh column of the table shows the expected additional raw achievement gain associated with Catholic school attendance. For example, HS&B Catholic school students, on average, gain 1.24 more points on the mathematics test between their sophomore and senior years than do HS&B public school students (i.e., 2.70 - 1.46 = 1.24). While the corresponding baseline Catholic school effect on achievement gains in mathematics is similar for NELS students (i.e., 5.48 - 4.37 = 1.11), this correspondence is misleading because there are many more items on the NELS tests.[16]

Hoffer et al. (1985) interpreted the Catholic school effect on achievement gains as grade-equivalent gains in achievement. I did not adopt this metric for presenting the regression models, but I included such calculations in Table A2 to provide a rough comparison of the raw test score differences between Hoffer et al.'s HS&B results and the NELS results in this article.

For Hoffer et al., the grade-equivalent gain measure of the Catholic school effect on achievement gains is simply the ratio of the raw Catholic school effect on achievement gains to half the average public school achievement gain (i.e., for HS&B mathematics 1.24/(.5*1.46) = 1.7). As shown in the last three columns of the table, when measured in grade-equivalent gains, the Catholic school effect on achievement gains appears to have declined substantially between the early 1980s and the early 1990s. It is hard to know how to interpret this measured decline, given that the compatibility of the tests is unknown even though both sets of tests were designed by the Educational Testing Service.

Nonetheless, the main story from Table A2 on which I focus in the article is consistency between the early 1980s and the early 1990s. In both the sophomore and senior years, Catholic school students score higher on tests of achievement in mathematics and reading than do public school students. In addition, Catholic school students increase their relative test performance advantage over public school students between the sophomore and senior years.

Consequences of Alternative Procedures for Handling Missing Data The coefficient estimates presented in Table 1 were estimat-

Table A1. Means and Standard Deviations of Variables Used in the Analysis

	NELS Source Variable	*Mean*	*SD*	*N*
Achievement Test Scores				
1990 Mathematics (IRT scaled)	F12XMIRR	44.23	13.48	13,593
1992 Mathematics (IRT scaled)	F22XMIRR	49.12	13.96	11,145
1990 Reading (IRT scaled)	F12XRIRR	30.90	9.84	13,608
1992 Reading (IRT scaled)	F22XRIRR	33.49	9.98	11,141
Catholic School Attendance				
Attended Catholic school in both the 10th and 12th grades	G10CTRL2, G12CTRL2	0.06	0.24	14,139
Attended Catholic school in the 8th grade	G8CTRL2	0.09	0.28	13,723
Family Background and Other Demographic Characteristics				
Socioeconomic Status	F2SES3	0.08	0.75	11,681
Urban	G10URBAN	0.26	0.44	14,139
Rural	G10URBAN	0.32	0.46	14,139
Northeast	G10REGO	0.19	0.39	14,139
North Central	G10REGO	0.27	0.44	14,139
South	G10REGO	0.34	0.47	14,139
Number of siblings	F1S90A-B	2.43	2.04	14,139
Have own bedroom	BYS35P, F1N21P	0.83	0.38	13,503
Two-parent family	F192A-F	0.77	0.42	14,061
Asian	F2RACE1	0.04	0.19	14,118
Hispanic	F2RACE1	0.09	0.29	14,118
Black	F2RACE1	0.12	0.32	14,118
American Indian	F2RACE1	0.01	0.10	14,118
Learning or physical disability	BYP47A-G	0.11	0.31	12,460
Educational Expectations and Parental Involvement				
Dad expects student to go to college	F1S48A	0.81	0.39	12,732
Mom expects student to go to college	F1S48B	0.81	0.39	11,960
Student plans to go to a 4-year college immediately after high school	F1S49, F1S51	0.52	0.50	13,696
Student expected to go to college in the 8th grade	BYS45	0.71	0.45	13,033
Parent volunteers at school	F1S106D	0.26	0.44	12,612
Time spent with parents	F1S44K	2.86	0.97	13,637
School Climate, School Track, and Individual Course Taking				
School climate	F1C95A-L	2.23	0.57	11,810
Academic track	F1HSPROG	0.39	0.49	12,896
Hours of homework per week	F1S36A1-2	7.25	5.09	13,507
Number of advanced mathematics courses taken	F2RAL1_C, F2RGEO_C, F2RAL2_C, F2RTRI_C	1.70	1.23	14,139
Number of advanced science courses taken	F2RBIO_C, F2RCHE_C F2RPHY_C	1.39	1.01	14,139
Number of semesters of English since the sophomore year	F1S24A, F2RENG_C	4.37	1.31	12,115

Note: Reference categories for race, urbanicity, and region are white non-Hispanic, suburban, and West. Data are weighted by the first follow-up questionnaire weight (F1QWT).

Table A2. Comparison of Differences in Raw Mean Test Scores between Public and Catholic High School Students

	Public School Students[a]			Catholic School Students[a]			Catholic School Effect on Achievement Gain (Sophomore to Senior Year)		
	Sophomores	Seniors	Achievement Gain	Sophomores	Seniors	Achievement Gain	Raw	*SE* of Raw	As Grade Equivalent
HS&B 1980 and 1982[b]									
Mathematics test	19.10	20.56	1.46	23.00	25.70	2.70	1.24	.19	1.7
Reading test	9.30	10.30	1.00	11.00	12.66	1.66	.66	.12	1.3
NELS 1990 and 1992[c]									
Mathematics test	44.58	48.95	4.37	49.25	54.73	5.48	1.11	.31	.51
Reading test	31.09	33.37	2.28	34.79	37.35	2.56	.28	.33	.25

[a]Hoffer, Greeley, and Coleman (1985) did not report a sample size for their results. For the NELS results, the number of public school students is 10,129 for the reading test and 10,121 for the mathematics test. The corresponding Ns for Catholic school students are 716 and 719.

[b]Table 2.1 of Hoffer et al. (1985) is the template for this table and the source of the HS&B results. However, Hoffer et al.'s Table 2.1 only contains the numbers presented in columns 1, 3, 7, 8, and 9 of this table. Although the senior-year public school score in column 2 and the Catholic school difference in column 6 could be easily calculated from their Table 2.1, the sophomore- and senior-year Catholic school scores in columns 4 and 5 are only best guesses based on a back-of-the-envelope calculation after comparing Table 2.1 with the numbers reported in Greeley's separate analysis in Table 1.1 of Hoffer et al.

[c]The NELS tests were scaled with item response theory (IRT). The test scores used here and elsewhere in the article, are the IRT-estimated number right scores. This is the most comparable available measure to the HS&B test score, which was an actual number right on the same test that was taken in both the sophomore and senior years.

ed from models that use best-subset regression imputation of missing data. Hoffer et al. (1985) instead used pairwise deletion of missing data, as in earlier rounds of their research (e.g., Coleman et al. 1982). Critics of their research argued that pairwise deletion may have artificially inflated estimates of the Catholic school effect (Noell 1982; Willms 1985). In retrospect, as I show later, it appears that such concerns were overemphasized.

In this appendix, three alternative sets of coefficient estimates are presented in Tables A3 and A4.[17] The first set of estimates (column 1 of Tables A3 and A4) is generated with the procedure adopted by Hoffer et al.—pairwise deletion of missing data. With this procedure, the covariance matrix of the variables that is used for regression estimation is calculated element by element, such that the covariance between each pair of variables is calculated for all students with nonmissing data on both variables. As a result, individuals only contribute to the portion of the covariance matrix for which they have valid data. Pairwise deletion of missing data is generally considered to be unacceptable in current research because the covariance matrix does not apply to any single identifiable sample, and this lack of correction between the sample and the covariance matrix results in both conceptual difficulties for the computation of standard errors (there is no single *N* for all coefficients) and related technical problems (the covariance matrix may not be invertible).

The second set of estimates (column 4 of Tables A3 and A4) is calculated using listwise deletion of missing data. With this procedure, individuals who have missing values on any of the variables used for any of the models are dropped from the estimation sample. Although listwise deletion of cases is the most common alternative to pairwise deletion, it is defensible only in rare circumstances, such as when the data are missing completely at random.

The third set of estimates (column 7 of Tables A3 and A4) is calculated using best-subset regression imputation for missing values. With this procedure, missing values for the independent variables of a regression model are imputed as a function of whatever nonmissing data are available on the other independent variables in the model. All individuals with at least some nonmissing data are therefore included in the model. Among the three missing data procedures utilized here, best-subset regression imputation is most easily defendable.[18]

As shown in Tables A3 and A4, regression estimates from models produced with pairwise deletion are similar to those produced with best-subset regression imputation. From this similarity, I conclude (with implicit reference to the similarity of the HS&B and NELS surveys) that there is little reason to question the estimates of Coleman and his colleagues simply because they used pairwise deletion of missing data. Moreover, because the estimates from models produced with listwise deletion are substantially different from those produced by either of the other methods, there may be reason to question the research on the Catholic school effect that embraces alternative listwise deletion procedures (e.g., Noell 1982).

Although I do not present the results here, it can be shown that students from disadvantaged families and students with low test scores are more likely to have missing data. Since these students are also more likely to attend public schools, listwise deletion artificially decreases estimates of the Catholic school effect because dropping respondents with any missing data increases the average test scores of public school students relative to the average test scores of Catholic school students.

Because I regard the best-subset regression imputation models as the models that use the most reliable procedure for handling missing data, I confine the discussion in the text to the regression estimates of the Catholic school effects that are presented in column 7 of Tables A3 and A4. However, one should note that the estimates from the pairwise deletion models, which most closely correspond to Hoffer et al.'s models, are similar to the estimates from the best-subset regression imputation models.

Table A3. OLS Regression Estimates of the Catholic School Effect on 12th-Grade Achievement (Dependent Variable: 1992 IRT-Scaled Test Score)

	Pairwise Deletion of Missing Data			Listwise Deletion of Missing Data			Best-subset Regression Imputation of Missing Data		
	Catholic School Coefficient	*SE*	*N*[a]	Catholic School Coefficient	*SE*	*N*	Catholic School Coefficient	*SE*	*N*
Mathematics Achievement Models[b]									
1. Catholic school dummy variable only	5.78	.56	11,145	3.53	.99	5,150	5.78	.84	10,835
2. 1 + family background and demographics	2.72	.60	11,145	1.12	.93	5,150	2.54	.79	10,835
3. 2 + expectations and parental involvement	1.25	.57	11,145	.39	.96	5,150	1.14	.78	10,835
4. 3 if course-taking variables not missing	1.45	.62	11,145	.46	.98	4,867	1.36	.79	10,003
5. 4 + climate, track, and course taking	-1.35	.57	11,145	-1.41	.90	4,867	-1.44	.75	10,003
6. 5 + attended Catholic school in the 8th grade	-1.92	.77	11,145	-1.60	1.36	4,867	-1.98	1.03	10,003
Reading Achievement Models[b]									
1. Catholic school dummy variable only	3.98	.40	11,141	2.59	.65	5,149	3.98	.57	10,840
2. 1 + family background and demographics	1.77	.44	11,141	.98	.66	5,149	1.67	.52	10,840
3. 2 + expectations and parental involvement	1.16	.44	11,141	.70	.68	5,149	1.08	.51	10,840
4. 3 if course-taking variables not missing	1.08	.47	11,141	.55	.67	4,865	1.01	.50	10,003
5. 4 + climate, track, and course taking	.03	.48	11,141	-.13	.67	4,865	-.05	.55	10,003
6. 5 + attended Catholic school in the 8th grade	- .79	.65	11,141	-.73	1.08	4,865	-.85	.78	10,003

[a] *N* is for the dependent variable only.

[b] For all the models, the data are weighted by the first follow-up questionnaire weight. For the listwise deleted models and the imputed data models, the standard errors are robust standard errors, calculated with STATA's implementation of White's sandwich variance estimator modified to adjust further for clustering within schools.

Table A4. OLS Regression Estimates of the Catholic School Effect on Achievement Gains between the 10th and 12th Grades from Models with a Lagged Independent Variable for Prior 10th-Grade Achievement (Dependent Variable: 1992 IRT-Scaled Test Score - 1990 IRT-Scaled Test Score)

	Pairwise Deletion of Missing Data			Listwise Deletion of Missing Data			Best-subset Regression Imputation of Missing Data		
	Catholic School Coefficient	*SE*	*N*[a]	Catholic School Coefficient	*SE*	*N*	Catholic School Coefficient	*SE*	*N*
Mathematics Achievement Models[b]									
1. Catholic school dummy variable only	1.31	.22	10,835	.40	.40	5,150	1.31	.31	10,835
2. 1 + family background and demographics	1.02	.27	10,835	.65	.43	5,150	.99	.33	10,835
3. 2 + expectations and parental involvement	.73	.28	10,835	.44	.44	5,150	.71	.33	10,835
4. 3 if course-taking variables not missing	.88	.30	10,003	.41	.46	4,867	.86	.34	10,003
5. 4 + climate, track, and course taking	.25	.31	10,003	-.10	.46	4,867	.23	.36	10,003
6. 5 + attended Catholic school in the 8th grade	-.34	.42	10,003	-.56	.68	4,867	-.34	.52	10,003
Reading Achievement Models[b]									
1. Catholic school dummy variable only	.91	.23	10,840	.99	.36	5,149	.91	.30	10,840
2. 1 + family background and demographics	.52	.28	10,840	.38	.39	5,149	.50	.32	10,840
3. 2 + expectations and parental involvement	.47	.29	10,840	.35	.39	5,149	.44	.32	10,840
4. 3 if course-taking variables not missing	.65	.31	10,003	.21	.37	4,865	.63	.30	10,003
5. 4 + climate, track, and course taking	.44	.34	10,003	-.05	.40	4,865	.40	.34	10,003
6. 5 + attended Catholic school in the 8th grade	.14	.46	10,003	-.36	.62	4,865	.10	.48	10,003

[a] *N* is for the dependent variable only.

[b] For all the models, the data are weighted by the first follow-up questionnaire weight. For the listwise deleted models and the imputed data models, the standard errors are robust standard errors, calculated with STATA's implementation of White's sandwich variance estimator modified to adjust further for clustering within schools.

APPENDIX B

Connections with the Econometric Treatment Self-Selection Model

The fundamental principles of the counterfactual model of causality are closely related (indeed, by some accounts, exactly equivalent) to those of econometric self-selection bias models. For readers who are familiar with the econometric literature, this appendix demonstrates the connections between these complementary frameworks.

Consider first the usual characterization of generic omitted variable bias with reference to Equations 1 and 2. If the "long" regression in Equation 2 is the true model, then the "short" regression in Equation 1 is misspecified. As a result, estimates of the coefficient d from Equation 1 will be biased and inconsistent. Traditionally, the bias is described by first noting that the error term of Equation 1 collectively represents all omitted variables that determine the scores on the 12th-grade test (including idiosyncratic individual-level determinants and random measurement error). Thus, for the model in Equation 1, the variables X_1 through X_k are omitted variables that are embedded in its error term. If, on average, there are differences between Catholic school students and public school students on the family background variables X_1 through X_k, as is always shown to be the case in empirical research, then for Equation 1, the dummy variable *Cath* is correlated with the error term e, and the standard OLS regression estimator of d yields biased and inconsistent estimates.

Suppose that we now have a data set $\{Y_i, C_i, \mathbf{X}_i, \mathbf{Z}_i\}_{i=1}^{n}$, where Y_i is an observed test score, C_i is a dummy variable for Catholic school attendance, $\mathbf{X}_i$ is a 1 by k row vector of the variables X_1 through X_k of determinants of learning (which may also be determinants of school sector decisions), and $\mathbf{Z}_i$ is a row vector of exogenous variables that determine school sector decisions but do not determine learning. Thus, we now have access to a set of observed variables in $\mathbf{Z}_i$ that predict sector selection, and we also think of some of our learning determinants in $\mathbf{X}_i$ as predictors of sector selection.

The most common representation of the selection bias model is the linear two-equation selection model for treatment evaluation (see Heckman and Robb 1985). For a simplified (and pure) version of this model, first write the potential outcomes introduced in the main text of the article as deviations from their unconditional population means: $Y_i^c = \bar{Y}^c + u_i^c$ and $Y_i^p = \bar{Y}^p + u_i^p$. Then, substitute these expressions into the observation rule in Equation 6 and re-arrange terms to write the observed Y_i as

$$\text{(B1)}\quad Y_i = \bar{Y}^p + C_i(\bar{Y}^c - \bar{Y}^p) + \{u_i^p + C_i(u_i^c - u_i^p)\},$$

where the term in braces is analogous to a conventional error term and represents individual-level departures from the relevant population means of the potential outcomes.

The selection model is then specified by introducing a latent continuous variable for selection into the Catholic school sector:

$$\text{(B2)}\quad C_i^* = \mathbf{S}_i\alpha + v_i$$

where $C_i = 1$ if $C_i^* \geq 0$, $C_i = 0$ if $C_i^* < 0$, and $\mathbf{S}_i$ is a row vector of variables that determine selection into the Catholic school sector (and therefore includes all elements of $\mathbf{Z}_i$ and possibly some elements of $\mathbf{X}_i$) through effects summarized by the parameter vector α. The mean-zero error term, v_i, represents for each individual the combined effects of unobserved and completely random effects on school sector selection. The explicit specification of a separable error term, v_i, in Equation B2 is the most important feature that distinguishes the selection equation from the propensity score defined in Equation 13.

If some of the variables in $\mathbf{X}_i$ are included in $\mathbf{S}_i$, then there is said to be "selection on the observables" because the index function $\mathbf{S}_i$ will be correlated with the error term in the braces in Equation B1. If, in addition, v_i, is purely idiosyncratic, or otherwise mean independent of the error term in braces in Equation B1, then an analysis of covariance model or a propensity score matching estimator (if properly specified) can be used to eliminate the selection bias.

If some variables that are embedded in v_i can also be considered predictors of learning (i.e., additional unobserved variables $X_{k+1}...X_l$) or if accurately perceived potential learning gains (i.e., Y_i^c - Y_i^p) determine sector selection, then there is said to be "selection on the unobservables" because v_i will be correlated with the error term in braces in Equation B1. In this case, a perfect implementation of analysis of covariance or propensity-score matching methods with available data will be able to recover only an estimate of the causal effect of Catholic schooling for those who typically attend Catholic schools, as in Equations 11 and 12.

NOTES

1. See Lieberson (1985) for one of the most cogent and extended criticisms of this general research strategy.

2. This is only one particular way in which the lagged variable regression model can go wrong. As Winship and Morgan (1999:698–700) showed, there are circumstances in which it is likely that estimates of the lag coefficient *l* are too large and thus that the estimation of Equation 3 will instead underestimate the true treatment effect. Indeed, the only strong justification for the use of Equation 3 to estimate *d* requires (among other things, including parallel linear regressions) that *Test10* be a true pretest and for attendance at a Catholic school in the 11th and 12th grades to be a function of *Test10* (see, Holland and Rubin 1983; Rubin 1977). In this case, the analysis of covariance correctly adjusts for regression to the mean.

3. When the regression coefficients on X_1 through X_k are permitted to vary across school sectors by estimating sector-specific regression models, E1, E2, and E3 are calculated by plugging the mean values of X_1 through X_k (respectively, for the whole sample, for the sample of public school students, and for the sample of Catholic school students) into the sector-specific regression models. See Murnane, Newstead, and Olsen (1985) for a further discussion.

4. See Sobel (1996) and Winship and Morgan (1999) for reviews of this literature written for sociologists.

5. See Heckman, Smith, and Clements (1997) for a general discussion of causal effect heterogeneity.

6. For researchers who are familiar with the literature on econometric selection bias, which addresses these issues in a slightly different manner (as reviewed for sociologists in Winship and Mare 1992), Appendix B provides the explicit connections between that framework and the ideas presented in this section. In current research, especially in the work of Heckman and his colleagues (e.g., Heckman, LaLonde, and Smith 1999), the two frameworks are the same.

7. Notice that through omission in the argument that follows, I have implicitly ruled out the possibility that conditioning on pretreatment variables would allow us to assert that $(\bar{Y}^c_{i\in P}|\mathbf{S}_i) = (\bar{Y}_i|C_i = 1,\mathbf{S}_i)$. I have therefore assumed that the Catholic school effect varies across the population partition, such that students who are enrolled in Catholic schools would be more likely than those who are not to benefit from Catholic schooling. Even more subtly, I have also implicitly assumed that those students who attend Catholic schools would not perform substantially worse than observationally equivalent public school students if they were instead educated in public schools. These auxiliary assumptions are based on a notion of behavioral self-interest, would be categorized as positive self-selection on unobservables in only one direction (in the framework of Appendix B), and are certainly worthy of extended evaluation.

8. Regression methods attempt to solve a related curse of dimensionality through model-based extrapolation across the dimensions of $\mathbf{S}_i$. Unfortunately, when the true model of learning is not known, regression-based extrapolation can go wrong, as is shown in many places, especially well for such models as Equations 2 and 3 in Holland and Rubin (1983).

9. To their credit, Hoffer et al. (1985) used a rough propensity-score technique in a latter portion of their article, stratifying the entire HS&B sample into quintiles of the propensity score. The inexactness of this match surely contributes to the fluctuation in their estimates across the quintiles. Thus, to give credit where it is due, the matching estimates I offer here can be read as a refinement of a

technique first adopted by Coleman and his colleagues (even before the methodological literature on propensity score matching had fully developed).

10. In more ambitious propensity-score approaches in which the ignorability of treatment assignment is invoked, it is sometimes assumed that students with the same propensity score are completely exchangeable (i.e., equivalent in all respects relevant for estimation of the Catholic school effect for all students). See note 11 for a further discussion of this issue.

11. If treatment assignment is ignorable (see Rosenbaum 1984b; Rosenbaum and Rubin 1983, 1985)—an assumption I do not invoke here but that is prominent in the early matching literature—one can extend the inferential reach of the propensity-score matching estimator by removing the conditioning on inclusion in the set C on the left side of Estimator M2. In that case, Estimator M2 would yield a consistent estimate of the average causal effect of Catholic schooling in Equation 7 (subject to the usual maintenance of SUTVA).

12. I use the same linear specifications employed for the regression models. Given the criticism I offered for such specifications, this may seem to be disingenuous. Ideally, in sensitivity analysis mode, one would perform alternative adjustments with alternative specifications. Space constraints do not permit such an exhaustive analysis. Nonetheless, since the adjustments are performed within matched strata of respondents, specification bias is minimized.

13. It is not necessarily the case that a larger data set would yield closely matched public school students for all Catholic school students. It may be the case, as in the JTPA evaluation of Heckman et al. (1999), that there are regions of nonoverlapping support in the population of self-selecting treatment and control cases.

14. Since I focus in this article mostly on the 12th-grade scores of students who are likely to graduate from high school, I did not estimate the possible impact of differential dropout patterns and school transfer patterns on gains in test scores. Arguments in the literature on the potential biases produced by the changing compositions of the students in the two school sectors are contradictory and must, at some point in the future, be reconciled in further research. On the one hand, explicitly modeling dropout patterns may show that the Catholic school effect is typically underestimated because Catholic schools are more likely to retain their worst students. On the other hand, explicitly modeling both dropout patterns and school transfer patterns may show that the Catholic school effect is typically overestimated because public schools often absorb the students who would otherwise be seen as dropouts from Catholic schools if public schools did not exist for them to drop into.

15. The only variable that did not have an acceptable counterpart in NELS is "mother working or not working when respondent was young" (Hoffer et al. 1985:note 4), derived from HS&B BB037b-c.

16. The NELS test scores are only "estimated number right" scores because the tests were designed to be scaled with item response theory (see Rock and Pollack 1995 for details).

17. Because of space limitations, I not discuss the standard errors of the estimates in the following presentation. It should be noted, however, that the standard errors for the models with pairwise deletion are too small because they were estimated with SPSS and not corrected in any way to account for the clustering of students within schools. Because the concept of a standard error in an OLS model that uses pairwise deletion is problematic anyway (since there is no single *N* for the estimation sample), there seemed to be little value in using a design effect to adjust the standard errors, as Hoffer et al. did. However, for both the listwise deleted models and best-subset regression imputation models (and thus for the models presented in the article), the standard errors are robust standard errors, calculated with STATA's implementation of White's sandwich variance estimator, which is modified also to adjust for clustering within schools.

18. Multiple imputation methods are the most sophisticated alternative. In essence, a set of values are imputed for each missing value, and the variances of these imputed sets

are explicitly incorporated into the standard errors of resulting parameter estimates (see Little and Rubin 1987).

REFERENCES

Alexander, Karl L., and Aaron M. Pallas. 1983. "Private Schools and Public Policy: New Evidence on Cognitive Achievement in Public and Private Schools." *Sociology of Education* 56:170–82.

—. 1985. "School Sector and Cognitive Performance: When Is a Little a Little?" *Sociology of Education* 58:115–28.

Allison, Paul D. 1990. "Change Scores as Dependent Variables in Regression Analysis." *Sociological Methodology* 20:93–114.

Bryk, Anthony S., Valerie Lee, and Peter Holland. 1993. *Catholic Schools and the Common Good.* Cambridge. MA: Harvard University Press.

Bush, Robert R., and William K. Estes, eds. 1959. *Studies in Mathematical Learning Theory.* Stanford, CA: Stanford University Press.

Carbonaro, William J. 1998. "A Little Help from My Friend's Parents: Intergenerational Closure and Educational Outcomes." *Sociology of Education* 71:295–313.

Chubb, John E., and Terry M. Moe. 1990. *Politics, Markets, and America's Schools.* Washington, DC: Brookings Institution.

Coleman, James S. 1964. *Introduction to Mathematical Sociology.* New York: Free Press.

—. 1981. *Longitudinal Data Analysis.* New York: Basic Books.

Coleman, James S., and Thomas Hoffer. 1987. *Public and Private Schools: The Impact of Communities.* New York: Basic Books.

Coleman, James S., Thomas Hoffer, and Sally Kilgore. 1982. *High School Achievement: Public, Catholic, and Private Schools Compared.* New York: Basic Books.

Figlio, David N., and Joe A. Stone. 1997. "School Choice and Student Performance: Are Private Schools Really Better?" (Discussion Paper 1141-97). Madison: Institute for Research on Poverty, University of Wisconsin.

Goldberger, Arthur S., and Glen G. Cain. 1982. "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report." *Sociology of Education* 55:103–22.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64:605–54.

—. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65:261-94.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith 1999. "The Economics and Econometrics of Active Labor Market Programs." Pp. 1865–2097 in *Handbook of Labor Economics, Volume III,* edited by Orley Ashenfelter and David Card. New York: Elsevier.

Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156–245 in *Longitudinal Analysis of Labor Market Data,* edited by James J. Heckman and Burton Singer. Cambridge, England: Cambridge University Press.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64:487–535.

Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. 1985. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 58:74–97.

Holland, Peter W., and Donald B. Rubin. 1983. "On Lord's Paradox." Pp 3–25 in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord,* edited by H. Wainer and S. Messick. Hillsdale, NJ: Lawrence Erlbaum.

Hoxby, Caroline M. 1996. "The Effects of Private School Vouchers on Schools and Students." Pp. 177–208 in *Holding Schools Accountable: Performance-Based Reform in Education,* edited by Helen F. Ladd. Washington, DC: Brookings Institution.

Judd, Charles M., and David A. Kenny. 1981. *Estimating the Effects of Social Interventions.* New York: Cambridge University Press.

Ladd, Helen F., ed. 1996. *Holding Schools Accountable: Performance-Based Reform in Education.* Washington, DC: Brookings Institution.

Lee, Valerie E., and Julia B. Smith. 1993. "Effects of School Restructuring on the Achievement of Middle-grade Students." *Sociology of Education* 66:164–87.

—. 1995. "Effects of High School Restructuring and Size on Early Gains in Achievement and Engagement." *Sociology of Education* 68:241-70.

Lee, Valerie E., Julia B. Smith, and Robert G. Croninger. 1997. "How High School Organization Influences the Equitable

Distribution of Learning in Mathematics and Science." *Sociology of Education* 70:128–50.
Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory.* Berkeley: University of California Press.
Little, Roderick, and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.
Morgan, Stephen L., and Aage B. Sørensen. 1999. "Parental Networks, Social Closure, and Mathematics Learning: A Test of James Coleman's Social Capital Explanation of School Effects." *American Sociological Review* 64:661–81.
Morgan, William R. 1983. "Learning and Student Life Quality of Public and Private School Youth." *Sociology of Education* 56:187–202.
Murnane, Richard J., Stephen E. Newstead, and Randall J. Olsen. 1985. "Comparing Public and Private Schools: The Puzzling Role of Selectivity Bias." *Journal of Business and Economic Statistics* 3:23–35.
National Center for Education Statistics. 1996. *National Education Longitudinal Study: 1988–94* [CD-ROM]. U.S. Department of Education, Office of Educational Research and Improvement [producer and distributor].
Neal, Derek. 1997. "The Effects of Catholic Secondary Schooling on Educational Achievement." *Journal of Labor Economics* 14:98–123.
Noell, Jay. 1982. "Public and Catholic Schools: A Reanalysis of "Public and Private Schools"." *Sociology of Education* 55:123–32.
Petersen, Paul E., and Bryan C. Hassel, eds. 1998. *Learning from School Choice.* Washington, DC: Brookings Institution.
Rasell, Edith, and Richard Rothstein, eds. 1993. *School Choice: Examining the Evidence.* Washington, DC: Economic Policy Institute.
Rock, Donald A., and Judith M. Pollack. 1995. *Psychometric Report for the NELS:88 Base Year (1988) through Second Follow-Up (1992).* Washington, DC: National Center for Education Statistics.
Rosenbaum, Paul R. 1984a. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147:656–66.
—. 1984b. "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment." *Journal of the American Statistical Association* 79:41–48.
Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
—. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods." *American Statistician* 39:33–38.
Rubin, Donald B. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2:1–26.
Rubin, Donald B., and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52:249–64.
—. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573–85.
Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27:325–53.
Sobel, Michael E. 1996. "An Introduction to Causal Inference." *Sociological Methods & Research* 24:353–79.
Sørensen, Aage B. 1996. "Educational Opportunities and School Effects." Pp. 207–25 in *James S. Coleman*, edited by J. Clark. London: Falmer.
Sørensen, Aage B., and Maureen T. Hallinan. 1977. "A Reconceptualization of School Effects." *Sociology of Education* 50:522–35.
Sørensen, Aage B., and Stephen L. Morgan. 2000. "School Effects: Theoretical and Methodological Issues." Pp. 137–60 in *Handbook of the Sociology of Education,* edited by Maureen T. Hallinan. New York: Kluwer.
Willms, J. Douglas. 1985. "Catholic-School Effects on Academic achievement: New Evidence from the High School and Beyond Follow-Up Study." *Sociology of Education* 58:98–114.
Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–50.
Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659–707.

***Stephen L. Morgan, Ph.D.**, is Assistant Professor, Department of Sociology, Cornell University, Ithaca, New York. His main fields of interest are sociology of education, social stratification, and quantitative methodology. He is currently studying black-white differences in the expectations and attainment relationship, developing new methods for predicting college entry and persistence using a stochastic decision tree model of commitment, and estimating rent-based positional advantages in the labor market, focusing on social-class differences.*

The author is grateful for the research assistance of Lisa F. Chavez; the programming efforts of Cheri Minton; the comments of Ken Frank, Felix Elwert, and David Harding; and discussions with Chris Winship, Aage Sørensen, and Stanley Lieberson. Financial support for this research was provided by the American Educational Research Association, which receives funds for its AERA Grants Program from the National Science Foundation and the National Center for Education Statistics (U.S. Department of Education) under NSF grant RED-9452861. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies. *Address all correspondence to Dr. Stephen L. Morgan, Department of Sociology, Cornell University, 366 Uris Hall, Ithaca, NY 14853; e-mail: slm45@cornell.edu.*