Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City

John BARNARD, Constantine E. FRANGAKIS, Jennifer L. HILL, and Donald B. RUBIN

The precarious state of the educational system in the inner cities of the United States, as well as its potential causes and solutions, have been popular topics of debate in recent years. Part of the difficulty in resolving this debate is the lack of solid empirical evidence regarding the true impact of educational initiatives. The efficacy of so-called "school choice" programs has been a particularly contentious issue. A current multimillion dollar program, the School Choice Scholarship Foundation Program in New York, randomized the distribution of vouchers in an attempt to shed some light on this issue. This is an important time for school choice, because on June 27, 2002 the U.S. Supreme Court upheld the constitutionality of a voucher program in Cleveland that provides scholarships both to secular and religious private schools. Although this study benefits immensely from a randomized design, it suffers from complications common to such research with human subjects: noncompliance with assigned "treatments" and missing data. Recent work has revealed threats to valid estimates of experimental effects that exist in the presence of noncompliance and missing data, even when the goal is to estimate simple intention-to-treat effects. Our goal was to create a better solution when faced with both noncompliance and missing data. This article presents a model that accommodates these complications that is based on the general framework of "principal stratification" and thus relies on more plausible assumptions than standard methodology. Our analyses revealed positive effects on math scores for children who applied to the program from certain types of schools—those with average test scores below the citywide median. Among these children, the effects are stronger for children who applied in the first grade and for African-American children.

KEY WORDS: Causal inference; Missing data; Noncompliance; Pattern mixture models; Principal stratification; Rubin causal model; School choice.

1. INTRODUCTION

There appears to be a crisis in America's public schools. "More than half of 4th and 8th graders fail to reach the most minimal standard on national tests in reading, math, and science, meaning that they probably have difficulty doing grade-level work" (Education Week 1998). The problem is worse in high poverty urban areas. For instance, although only 43% of urban fourth-graders achieved a basic level of skill on a National Assessment of Educational Progress (NAEP) reading test, a meager 23% of those in high-poverty urban schools met this standard.

One of the most complicated and contentious of educational reforms currently being proposed is school choice. Debates about the equity and potential efficacy of school choice have increased in intensity over the past few years. Authors making a case for school choice include Cobb (1992), Brandl (1998), and Coulson (1999). A collection of essays that report mainly positive school choice effects has been published by Peterson and Hassel (1998). Recent critiques of school choice include those by the Carnegie Foundation for the Advancement of Teaching (1992), Cookson (1994), Fuller and Elmore (1996), and Levin (1998).

In this article we evaluate a randomized experiment conducted in New York City made possible by the privately-funded School Choice Scholarships Foundation (SCSF). The SCSF program provided the first opportunity to examine the question of the potential for improved school performance (as well as parental satisfaction and involvement, school mobility, and racial integration) in private schools versus public schools using a carefully designed and monitored randomized field experiment. Earlier studies were observational in nature and thus subject to selection bias (i.e., nonignorable treatment assignment). Studies finding positive educational benefits from attending private schools include those of Coleman, Hoffer, and Kilgore (1982), Chubb and Moe (1990), and Derek (1997). Critiques of these studies include those of Goldberger and Cain (1982) and Wilms (1985). On June 27, 2002, the U.S. Supreme Court upheld the constitutionality of a voucher program in Cleveland that provides scholarships both to secular and religious private schools.

As occurs in most research involving human subjects, however, our study, although carefully implemented, suffered from complications due to missing background and outcome data and also to noncompliance with the randomly assigned treatment. We focus on describing and addressing these complications in our study using a Bayesian approach with the framework of principal stratification (Frangakis and Rubin 2002).

We describe the study in Section 2 and summarize its data complications in Section 3. In Section 4 we place the study in

John Barnard is Senior Research Statistician, deCODE Genetics, Waltham, MA-02451 (E-mail: *john.barnard@decode.is*). Constantine E. Frangakis is Assistant Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *cfrangak@jhsph.edu*). Jennifer L. Hill is Assistant Professor, School of International and Public Affairs, Columbia University, New York, NY 10027 (E-mail: *jh1030@columbia.edu*). Donald B. Rubin is John L. Loeb Professor of Statistics and Chair, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *rubin@stat.harvard.edu*). The authors thank the editor, an associate editor, and three reviewers for very helpful comments; David Myers and Paul E. Peterson as principal coinvestigators for this evaluation; and the School Choice Scholarships Foundation (SCSF) for cooperating fully with this evaluation. The work was supported in part by National Institutes of Health (NIH) grant RO1 EY 014314-01, National Science Foundation grants SBR 9709359 and DMS 9705158; and by grants from the following foundations: Achelis Foundation, Bodman Foundation, Lynde and Harry Bradley Foundation, Donner Foundation, Milton and Rose D. Friedman Foundation, John M. Olin Foundation, David and Lucile Packard Foundation, Smith Richardson Foundation, and the Spencer Foundation. The authors also thank Kristin Kearns Jordan and other members of the SCSF staff for their cooperation and assistance with data collection, and Daniel Mayer and Julia Kim, from Mathematica Policy Research, for preparing the survey and test score data and answering questions about that data. The methodology, analyses of data, reported findings and interpretations of findings are the sole responsibility of the authors and are not subject to the approval of SCSF or of any foundation providing support for this research.

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Applications and Case Studies
DOI 10.1198/016214503000071

the context of broken randomized experiments, a phrase apparently first coined by Barnard, Du, Hill, and Rubin (1998). We discuss the framework that we use in Section 5. We present our model's structural assumptions in Section 6, and its parametric assumptions in Section 7. We give the main results of the analysis in Section 8 (with some supplementary results in the Appendix). We discuss model building and checking in Section 9, and conclude the article in Section 10.

2. THE SCHOOL CHOICE SCHOLARSHIPS FOUNDATION PROGRAM

In February 1997 the SCSF announced that it would provide 1,300 scholarships to private school to "eligible" low-income families. Eligibility required that the children, at the time of application, be living in New York City, entering grades 1–5, currently attending a public school, and from families with incomes low enough to qualify for free school lunch. That spring, SCSF received applications from more than 20,000 students. To participate in the lottery that would award the scholarships, a family had to attend a session during which (1) their eligibility was confirmed, (2) a questionnaire of background characteristics was administered to the parents/guardians, and (3) a pretest was administered to the eligible children. The final lottery, held in May 1997, was administered by Mathematica Policy Research (MPR), and the SCSF offered winning families help in finding placements in private schools.

Details of the design have been described by Hill, Rubin, and Thomas (2000). The final sample sizes of children are displayed in Table 1. PMPD refers to the randomized design developed for this study (propensity matched pairs design). This design relies on propensity score matching (Rosenbaum and Rubin 1983), which was used to choose a control group for the families in the first application period, where there were more applicants that did not win the scholarship than could be followed. The "single" and "multi" classifications describe families that have one child and more than one child participating in the program.

Table 1. Sample Sizes in the SCSF Program

Family size	Treatment	PMPD	Randomized block 1	2	3	4	Subtotal	Total
Single	Scholarship	353	72	65	82	104	323	676
	Control	353	72	65	82	104	323	676
Multi	Scholarship	147	44	27	31	75	177	324
	Control	147	27	23	33	54	137	284
	Total	1,000					960	1,960

For period 1 and single-child families, Table 2 (taken from Barnard, Frangakis, Hill, and Rubin 2002, table 1.6) compares the balance achieved on background variables with the PMPD and two other possible designs: a simple random sample (RS) and a stratified random sample (STRS) of the same size, from the pool of all potential matching subjects at period 1. For the STRS, the strata are the "applicant's school" (low/high), which indicates whether the applicant child originates from a school that had average test scores below (low) or above (high) the citywide median in the year of application. Measures of comparison in Table 2 are Z statistics between the randomized arms. Overall, the PMPD produces better balance in 15 of the 21 variables compared with the RS design. The PMPD's balance was better in 11 variables and worse in 9 variables (1 tie) compared with the STRS, although the gains are generally larger in the former case than in the latter case. The table also demonstrates balance for the application periods 2–5, which were part of a standard randomized block design in which the blocks were each of the four periods, cross-classified by family size and by applicant's school.

More generally, the entire experiment is a randomized design where the assignment probabilities are a function of the following design variables: period of application, applicant's school, family size (single child versus multichild), and the estimated propensity scores from the PMPD. (For additional information on the design, see Hill et al. 2000.) Next, we focus on the study's main data complications.

Table 2. Design Comparisons in Balance of Background Variables: Single-Child Families. The Numbers Are Z Statistics From Comparing Observed Values of Variables Between Assignments

Variable	Application period 1: Simple random sample	Stratified random sample	PMPD	Periods 2–5: Randomized block
Applicant's school (low/high)	−.98	0	.11	.21
Grade level	−1.63	.03	−.03	−.39
Pretest read score	−.38	.65	.48	−1.05
Pretest math score	−.51	1.17	.20	−1.37
African-American	1.80	1.68	1.59	1.74
Mother's education	.16	.14	.09	1.67
In special education	.31	1.66	−.17	.22
In gifted program	.42	−1.16	−.13	.75
English main language	−1.06	−.02	−1.03	−.44
AFDC	−.28	.49	.83	−1.57
Food stamps	−1.08	−.27	.94	−1.31
Mother works	−1.26	−.30	−1.18	.40
Educational expectations	.50	1.79	.57	.19
Children in household	−1.01	−1.75	.41	−1.02
Child born in U.S.	.49	.73	−1.40	−.69
Length of residence	.42	.71	.66	−.78
Father's work missing	1.09	.70	0	.16
Catholic religion	−1.84	−.19	−.74	−.80
Male	.88	1.22	.76	.53
Income	−.38	−.62	.74	−1.21
Age as of 4/97	−1.57	.18	−.47	−.87

3. DATA COMPLICATIONS

The data that we use include the design variables, the background survey collected at the verification sessions, pretest data, and posttest data collected the following spring. The test scores used as outcomes in our analyses are grade-normed national percentile rankings of reading and math scores from the Iowa Test of Basic Skills (ITBS). The ITBS was used because it is not the test administered in the New York City public school system, which reduces the possibility of teachers in schools with participating children "teaching to the test."

Attempts to reduce missingness of data included requiring attendance at the initial verification sessions and providing financial incentives to attend the follow-up testing sessions. Despite these attempts, missingness of background variables did occur before randomization. In principle, such missingness is also a covariate and so does not directly create imbalance of subjects between randomized arms, although it does create loss in efficiency when, as in this study, background covariates are used in the analysis. For example, for single-child families, depending on application period and background strata, 34%–51% of the children's pretest scores were missing at the time of design planning. Since then, MPR has processed and provided an additional 7% and 14% of the reading and mathematics pretest scores. These scores were not as balanced between arms as when choosing the design (see Table 2), although the difference was not statistically significant at conventional levels. Hence we used all available pretest scores in the final analysis that is conditional on these scores.

Outcomes were also incomplete. Among the observed outcomes in single-child families, the average (standard deviation) was 23.5 (22.5) percentile rankings for mathematics and 28.1 (23.9) percentile rankings for reading, and unadjusted comparisons between randomized arms do not point to a notable difference. However, in contrast to missingness of background variables, missingness of outcomes that occurs after randomization is not guaranteed to be balanced between the randomized arms. For example, depending on application period and background strata, 18%–27% of the children did not provide posttest scores, and during design periods 2–5, the response is higher among scholarship winners (80%) than among the other children (73%). Analyses that would be limited to complete cases for these variables and the variables used to calculate the propensity score would discard more than half of the units. Moreover, standard adjustments for outcome missingness ignore its potential interaction with the other complications and generally make implicit and unrealistic assumptions.

Another complication was noncompliance; attendance at a private school was not perfectly correlated with winning a scholarship. For example, for single-child families, and depending on application period and background strata, 20%–27% of children who won scholarships did not use them (scholarships were in the amount of $1,400, which generally did not fully cover tuition at a private school), and 6%–10% of children that did not win scholarships were sent to private schools nevertheless.

Two additional complications limit our analysis sample. First, no pretest scores were obtained for applicants in kindergarten, because these children most likely had never been exposed to a standardized test, hence considerable time would have been spent instructing them on how to take a test, and there was concern that separating such young children from their guardians in this new environment might lead to behavioral issues. Hence we focus our analyses on the children who applied in first grade or above. Second, we do not yet have complete compliance data for the multichild families. Consequently, all analyses reported in this article are further limited to results for the 1,050 single-child families who were in grades 1–4 at the time of the spring 1997 application process.

4. THE STUDY AS A BROKEN RANDOMIZED EXPERIMENT

The foregoing deviations from the study's protocol clarify that our experiment does not really randomize attendance, but rather randomizes the "encouragement," using financial support, to attend a private school rather than a public school. Moreover, as in most encouragement studies, interest here focuses not only on the effect of encouragement itself (which will depend on what percentage of people encouraged would actually participate if the voucher program were to be more broadly implemented), but also on the effect of the treatment being encouraged—here, attending private versus public schools. If there were perfect compliance, so that all those encouraged to attend private school actually did so and all those not so encouraged stayed in public school, then the effect being estimated typically would be attributed to private versus public school attendance, rather than simply to the encouragement.

We focus on defining and estimating two estimands: the intention-to-treat (ITT) effect, the effect of the randomized encouragement on all subjects; and the complier average causal effect (CACE), the effect of the randomized encouragement on all subjects who would comply with their treatment assignment no matter which assignment they would be given (here, the children who would have attended private school if they had won a scholarship and would not have attended had they not won a scholarship). These quantities are defined more formally in Section 8.

In recent years there has been substantial progress in the analysis of encouragement designs, based on building bridges between statistical and econometric approaches to causal inference. In particular, the widely accepted approach in statistics to formulating causal questions is in terms of "potential outcomes." Although this approach has roots dating back to Neyman and Fisher in the context of perfect randomized experiments (Neyman 1923; Rubin 1990), it is generally referred to as Rubin's causal model (Holland 1986) for work extending the framework to observational studies (Rubin 1974, 1977) and including modes of inference other than randomization-based—in particular, Bayesian (Rubin 1978a, 1990). In economics, the technique of instrumental variables, due to Tinbergen (1930) and Haavelmo (1943), has been a main tool of causal inference in the type of nonrandomized studies prevalent in that field. Angrist, Imbens, and Rubin (1996) showed how the approaches can be viewed as completely compatible, thereby clarifying and strengthening each approach. The result was an interpretation of the instrumental variables technology as a way to approach a randomized experiment that suffers from noncompliance, such as a randomized encouragement design.

In encouragement designs with compliance as the only partially uncontrolled factor, and where there are full outcome data, Imbens and Rubin (1997) extended the Bayesian approach to causal inference of Rubin (1978a) to handle simple randomized experiments with noncompliance, and Hirano, Imbens, Rubin, and Zhou (2000) further extended the approach to handle fully observed covariates.

In encouragement designs with more than one partially uncontrolled factor, as with noncompliance and missing outcomes in our study, defining and estimating treatment effects of interest becomes more challenging. Existing methods (e.g., Robins, Greenland, and Hu 1999) are designed for studies that differ from our study in the goals and the degree of control of the aforementioned factors. (For a detailed comparison of such frameworks, see Frangakis and Rubin 2002.) Frangakis and Rubin (1999) studied a more flexible framework for encouragement designs with both noncompliance to treatment and missing outcomes, and showed that for estimation of either the ITT effect or CACE, one cannot in general obtain valid estimates using standard ITT analyses (i.e., analyses that ignore data on compliance behavior) or standard IV analyses (i.e., those that ignore the interaction between compliance behavior and outcome missingness). They also provided consistent moment-based estimators that can estimate both ITT and CACE under assumptions more plausible than those underlying more standard methods.

Barnard et al. (1998) extended that template to allow for missing covariate and multivariate outcome values; they stopped short, however, of introducing specific methodology for this framework. We present a solution to a still more challenging situation in which we have a more complicated form of noncompliance—some children attend private school without receiving the monetary encouragement (thus receiving treatment without having been assigned to it). Under assumptions similar to those of Barnard et al. (1998), we next fully develop a Bayesian framework that yields valid estimates of quantities of interest and also properly accounts for our uncertainty about these quantities.

5. PRINCIPAL STRATIFICATION IN SCHOOL CHOICE AND ROLE FOR CAUSAL INFERENCE

To make explicit the assumptions necessary for valid causal inference in this study, we first introduce "potential outcomes" (see Rubin 1979; Holland 1986) for all of the posttreatment variables. Potential outcomes for any given variable represent the observable manifestations of this variable under each possible treatment assignment. In particular, if child i in our study ($i = 1, \ldots, n$) is to be assigned to treatment z (1 for private school and 0 for public school), we denote the following: $D_i(z)$ for the indicator equal to 1 if the child attends private school and 0 if the child attends public school; $Y_i(z)$ for the potential outcomes if the child were to take the tests; and $R_{y_i}(z)$ for the indicators equal to 1 if the child takes the tests. We denote by Z_i the indicator equal to 1 for the observed assignment to private school or not, and let $D_i = D_i(Z_i)$, $Y_i = Y_i(Z_i)$, and $R_{y_i} = R_{y_i}(Z_i)$ designate the actual type of school, the outcome to be recorded by the test, and whether or not the child takes the test under the observed assignment.

The notation for these and the remaining definitions of relevant variables in this study are summarized in Table 3. In our study the outcomes are reading and math test scores, so the dimension of $Y_i(z)$ equals two, although more generally our template allows for repeated measurements over time.

Our design variables are application period, low/high indicators for the applicant's school, grade level, and propensity score. In addition, corresponding to each set of these individual-specific random variables is a variable (vector or matrix), without subscript i, that refers to the set of these variables across all study participants. For example, Z is the vector of treatment assignments for all study participants with ith element Z_i, and X is the matrix of partially observed background variables with ith row X_i.

The variable C_i (see Table 3), the joint vector of treatment receipt under both treatment assignments, is particularly important. Specifically, C_i defines four strata of people: compliers, who take the treatment if so assigned and take control if so assigned; never takers, who never take the treatment no matter

Table 3. Notation for the ith Subject

Notation	*Specifics*	*General description*
Z_i	1 if i assigned treatment 0 if i assigned control	Binary indicator of treatment assignment
$D_i(z)$	1 if i receives treatment under assignment z 0 if i receives control under assignment z	Potential outcome formulation of treatment receipt
D_i	$D_i(Z_i)$	Binary indicator of treatment receipt
C_i	c if $D_i(0) = 0$ and $D_i(1) = 1$ n if $D_i(0) = 0$ and $D_i(1) = 0$ a if $D_i(0) = 1$ and $D_i(1) = 1$ d if $D_i(0) = 1$ and $D_i(1) = 0$	Compliance principal stratum: c = complier; n = never taker; a = always taker; d = defier
$Y_i(z)$	$(Y_i^{(math)}(z), Y_i^{(read)}(z))$	Potential outcomes for math and reading
Y_i	$(Y_i^{(math)}(Z_i), Y_i^{(read)}(Z_i))$	Math and reading outcomes under observed assignment
$Ry_i^{(math)}(z)$	1 if $Y_i^{(math)}(z)$ would be observed 0 if $Y_i^{(math)}(z)$ would not be observed	Response indicator for $Y_i^{(math)}(z)$ under assignment z; similarly for $Ry_i^{(read)}(z)$
$Ry_i(z)$	$(Ry_i^{(math)}(z), Ry_i^{(read)}(z))$	Vector of response indicators for $Y_i(z)$
Ry_i	$(Ry_i^{(math)}(Z_i), Ry_i^{(read)}(Z_i))$	Vector of response indicators for Y_i
W_i	$(W_{i1}, \ldots, W_{iK})$	Fully observed background and design variables
X_i	$(X_{i1}, \ldots, X_{iQ})$	Partially observed background variables
Rx_i	$(Rx_{i1}, \ldots, Rx_{iQ})$	Vector of response indicators for X_i

the assignment; always takers, who always take the treatment no matter the assignment; and defiers, who do the opposite of the assignment no matter its value. These strata are not fully observed, in contrast to observed strata of actual assignment and attendance (Z_i, D_i). For example, children who are observed to attend private school when winning the lottery are a mixture of compliers $(C_i = c)$ and always takers $(C_i = a)$. Such explicit stratification on C_i dates back at least to the work of Imbens and Rubin (1997) on randomized trials with noncompliance, was generalized and taxonomized by Frangakis and Rubin (2002) to posttreatment variables in partially controlled studies, and is termed a "principal stratification" based on the posttreatment variables.

The principal strata C_i have two important properties. First, they are not affected by the assigned treatment. Second, comparisons of potential outcomes under different assignments within principal strata, called principal effects, are well-defined causal effects (Frangakis and Rubin 2002). These properties make principal stratification a powerful framework for evaluation, because it allows us to explicitly define estimands that better represent the effect of attendance in the presence of noncompliance, and to explore richer and explicit sets of assumptions that allow estimation of these effects under more plausible than standard conditions. Sections 6 and 7 discuss such a set of more flexible assumptions and estimands.

6. STRUCTURAL ASSUMPTIONS

First, we state explicitly our structural assumptions about the data with regard to the causal process, the missing data mechanism and the noncompliance structure. These assumptions are expressed without reference to a particular parametric distribution and are the ones that make the estimands of interest identifiable, as also discussed in Section 6.4.

6.1 Stable Unit Treatment Value Assumption

A standard assumption made in causal analyses is the *stable unit treatment value assumption* (SUTVA), formalized with potential outcomes by Rubin (1978a, 1980, 1990). SUTVA combines the no-interference assumption (Cox 1958) that one unit's treatment assignment does not affect another unit's outcomes with the assumption that there are "no versions of treatments." For the no-interference assumption to hold, whether or not one family won a scholarship should not affect another family's outcomes, such as their choice to attend private school or their children's test scores. We expect our results to be robust to the types and degree of deviations from no interference that might be anticipated in this study. To satisfy the "no versions of treatments," we need to limit the definition of private and public schools to those participating in the experiment. Generalizability of the results to other schools must be judged separately.

6.2 Randomization

We assume that scholarships have been randomly assigned. This implies that

$$p(Z \mid Y(1), Y(0), X, W, C, Ry(0), Ry(1), Rx, \theta) = p(Z \mid W^*, \theta) = p(Z \mid W^*),$$

where W^* represents the design variables in W and θ is generic notation for the parameters governing the distribution of all the variables. There is no dependence on θ, because there are no unknown parameters governing the treatment-assignment mechanism. MPR assigned the scholarships by lottery, and the randomization probabilities within for applicant's school (low/high) and application period are known.

6.3 Noncompliance Process Assumptions: Monotonicity and Compound Exclusion

We assume monotonicity, that there are no "defiers"—that is, for all i, $D_i(1) \geq D_i(0)$ (Imbens and Angrist 1994). In the SCSF program, defiers would be families who would not use a scholarship if they won one, but would pay to go to private school if they did not win a scholarship. It seems implausible that such a group of people exists; therefore, the monotonicity assumption appears to be reasonable for our study.

By definition, the never takers and always takers will participate in the same treatment (control or treatment) regardless of which treatment they are randomly assigned. For this reason, and to facilitate estimation, we assume compound exclusion: The outcomes and missingness of outcomes for never takers and always takers are not affected by treatment assignment. This compound exclusion restriction (Frangakis and Rubin 1999) generalizes the standard exclusion restriction (Angrist et al. 1996; Imbens and Rubin 1997) and can be expressed formally for the distributions as

$$p(Y(1), Ry(1) \mid X^{\text{obs}}, Rx, W, C = n) = p(Y(0), Ry(0) \mid X^{\text{obs}}, Rx, W, C = n), \text{ for never takers}$$

and

$$p(Y(1), Ry(1) \mid X^{\text{obs}}, Rx, W, C = a) = p(Y(0), Ry(0) \mid X^{\text{obs}}, Rx, W, C = a), \text{ for always takers.}$$

Compound exclusion seems more plausible for never takers than for always takers in this study. Never takers stayed in the public school system whether they won a scholarship or not. Although a disappointment about winning a scholarship but still not being able to take advantage of it can exist, it is unlikely to cause notable differences in subsequent test scores or response behaviors.

Always takers, on the other hand, might have been in one private school had they won a scholarship or in another if they had not, particularly because those who won scholarships had access to resources to help them find an appropriate private school and had more money (up to $1,400) to use toward tuition. In addition, even if the child had attended the same private school under either treatment assignment, the extra $1,400 in family resources for those who won the scholarship could have had an effect on student outcomes. However, because in our application the estimated percentage of always takers is so small (approximately 9%)—an estimate that is robust, due to the randomization, to relaxing compound exclusion—there is reason to believe that this assumption will not have a substantial impact on the results.

Under the compound exclusion restriction, the ITT comparison of all outcomes $Y_i(0)$ versus $Y_i(1)$ includes the null comparison among the subgroups of never takers and always takers. Moreover, by monotonicity, the compliers $(C_i = c)$ are the only group of children who would attend private school if and only if

offered the scholarship. For this reason, we take the CACE (Imbens and Rubin 1997), defined as the comparison of outcomes $Y_i(0)$ versus $Y_i(1)$ among the principal stratum of compliers, to represent the effect of attending public versus private school.

6.4 Latent Ignorability

We assume that potential outcomes are independent of missingness given observed covariates conditional on the compliance strata, that is,

$$p(Ry(0), Ry(1) \mid Rx, Y(1), Y(0), X^{\text{obs}}, W, C, \theta) = p(Ry(0), Ry(1) \mid Rx, X^{\text{obs}}, W, C, \theta).$$

This assumption represents a form of latent ignorability (LI) (Frangakis and Rubin 1999) in that it conditions on variables that are (at least partially) unobserved or latent—here compliance principal strata C. We make it here first because it is more plausible than the assumption of standard ignorability (SI) (Rubin 1978a; Little and Rubin 1987), and second, because making it leads to different likelihood inferences.

LI is more plausible than SI to the extent that it provides a closer approximation to the missing-data mechanism. The intuition behind this assumption in our study is that for a subgroup of people with the same covariate missing-data patterns, Rx; similar values for covariates observed in that pattern, X^{obs}; and the same compliance stratum C, a flip of a coin could determine which of these individuals shows up for a posttest. This is a more reasonable assumption than SI, because it seems quite likely that for example, compliers, would exhibit different attendance behavior for posttests than, say, never takers (even conditional on other background variables). Explorations of raw data from this study across individuals with known compliance status provide empirical support that C is a strong factor in outcome missingness, even when other covariates are included in the model. This fact is also supported by the literature for noncompliance (see, e.g., The Coronary Drug Project Research Group 1980).

Regarding improved estimation, when LI (and the preceding structural assumptions) hold but the likelihood is constructed assuming SI, the underlying causal effects are identifiable (alternatively, the posterior distribution with increasing sample size converges to the truth) only if the additional assumption is made that within subclasses of subjects with similar observed variables, the partially missing compliance principal stratum C is not associated with potential outcomes. However, as noted earlier, this assumption is not plausible.

Theoretically, only the *structural* assumptions described earlier are needed to identify the underlying ITT and CACE causal effects (Frangakis and Rubin 1999). Estimation based solely on those identifiability relations in principle requires very large sample size and explicit conditioning (i.e., stratification) on the subclasses defined by the observed part of the covariates, X^{obs}, and the pattern of missingness of the covariates, Rx, as well as implicit conditioning (i.e., deconvolution of mixtures) on the sometimes missing compliance principal stratum C. This works in large samples because covariate missingness and compliance principal stratum are also covariates (i.e., defined pretreatment), so samples within subclasses defined by X^{obs}, Rx, and C themselves represent randomized experiments. With our experiment's sample size, however, performing completely separate analyses on all of these strata is not necessarily desirable or feasible. Therefore, we consider more parsimonious modeling approaches, which have the role of assisting, not creating, inference in the sense that results should be robust to different parametric specifications (Frangakis and Rubin 2001, rejoinder).

7. PARAMETRIC PATTERN MIXTURE MODEL

Generally speaking, constrained estimation of separate analyses within missing-data patterns is the motivation behind pattern mixture modeling. Various authors have taken pattern mixture model approaches to missing data, including Little (1993, 1996), Rubin (1978b), and Glynn, Laird, and Rubin (1993). Typically, pattern mixture models partition the data with respect to the missingness of the variables. Here we partition the data with respect to the covariate missing-data patterns Rx, as well as compliance principal strata C, design variables W, and the main covariates X^{obs}. This represents a *partial* pattern mixture approach. One argument in favor of this approach is that it focuses the model on the quantities of interest in such a way that parametric specifications for the marginal distributions of Rx, W, and X^{obs} can be ignored. To capitalize on the structural assumptions, consider the factorization of the joint distribution for the potential outcomes and compliance strata conditional on the covariates and their missing-data patterns,

$$\begin{aligned} p\big(Y_i(0), Y_i(1), Ry_i(0), Ry_i(1), C_i \mid W_i, X_i^{\text{obs}}, Rx_i, \theta\big) &= p\big(C_i \mid W_i, X_i^{\text{obs}}, Rx_i, \theta^{(\text{C})}\big) \\ &\quad \times p\big(Ry_i(0), Ry_i(1) \mid W_i, X_i^{\text{obs}}, Rx_i, C_i, \theta^{(\text{R})}\big) \\ &\quad \times p\big(Y_i(0), Y_i(1) \mid W_i, X_i^{\text{obs}}, Rx_i, C_i, \theta^{(\text{Y})}\big), \end{aligned}$$

where the product in the last line follows by latent ignorability and $\theta = (\theta^{(\text{C})}, \theta^{(\text{R})}, \theta^{(\text{Y})})'$. Note that the response pattern of covariates for each individual is itself a covariate. The parametric specifications for each of these components are described next.

7.1 Compliance Principal Stratum Submodel

The compliance status model contains two conditional probit models, defined using indicator variables $C_i(c)$ and $C_i(n)$, for whether individual i is a complier or a never taker:

$$C_i(n) = 1 \text{ if } C_i(n)^* \equiv g_1(W_i, X_i^{\text{obs}}, Rx_i)'\beta^{(\text{C},1)} + V_i \le 0$$

and

$$C_i(c) = 1 \text{ if } C_i(n)^* > 0 \quad \text{and} \quad C_i(c)^* \equiv g_0(W_i, X_i^{\text{obs}}, Rx_i)'\beta^{(\text{C},2)} + U_i \le 0,$$

where $V_i \sim \text{N}(0, 1)$ and $U_i \sim \text{N}(0, 1)$ independently.

The link functions, g_0 and g_1, attempt to strike a balance between on the one hand including all of the design variables as well as the variables regarded as most important either in predicting compliance or in having interactions with the treatment effect and on the other hand maintaining parsimony. The results discussed in Section 8 use a compliance component model whose link function, g_1, is linear in, and fits distinct parameters for, an intercept, applicant's school (low/high), indicators for application period, propensity scores for subjects applying in

the PMPD period and propensity scores for the other periods, indicators for grade of the student, ethnicity (1 if the child or guardian identifies herself as African-American, 0 otherwise), an indicator for whether or not the pretest scores of reading and math were available, and the pretest scores (reading and math) for the subjects with available scores. A propensity score for the students not in the PMPD period is not necessary from the design standpoint, and is in fact constant. However, to increase efficiency, and to reduce bias due to missing outcomes, here we include an "estimated propensity score value" for these periods, calculated as the function derived for the propensity score for students in the PMPD period and evaluated at the covariate values for the students in the other periods as well. Also, the foregoing link function g_0 is the same as g_1 except that it excludes the indicators for application period as well as the propensity scores for applicants who did not apply in the PMPD period (i.e., those for whom propensity score was not a design variable). This link function, a more parsimonious version of one we used in earlier models, was more appropriate to fit the relatively small proportion of always takers in this study. Finally, because the pretests were either jointly observed or jointly missing, one indicator for missingness of pretest scores is sufficient.

The prior distributions for the compliance submodel are $\beta^{(\mathrm{C},1)} \sim \mathrm{N}(\beta_0^{(\mathrm{C},1)}, \{\sigma^{(\mathrm{C},1)}\}^2\mathbf{I})$ and $\beta^{(\mathrm{C},2)} \sim \mathrm{N}(0, \{\sigma^{(\mathrm{C},2)}\}^2\mathbf{I})$ independently, where $(\sigma^{(\mathrm{C},1)})^2$ and $(\sigma^{(\mathrm{C},2)})^2$ are hyperparameters set at 10 and $\beta_0^{(\mathrm{C},1)}$ is a vector of 0s with the exception of the first element, which is set equal to $-\Phi^{-1}(1/3) * \{\sigma^{(\mathrm{C},1)} 1/n \sum_i^n (g'_{1,i} g_{1,i}) + 1\}^{1/2}$, where $g_{1,i} = g_1(W_i, X_i^{\mathrm{obs}}, Rx_i)$ and n is our sample size. These priors reflect our a priori ignorance about the probability that any individual belongs to each compliance status by approximately setting each of their prior probabilities at 1/3.

7.2 Outcome Submodel

We first describe the marginal distribution for the math outcome $Y^{(\mathrm{math})}$. To address the pile-up of many scores of 0, we posit the censored model

$$Y_i^{(\mathrm{math})}(z) = \begin{cases} 0 & \text{if } Y_i^{(\mathrm{math}),*}(z) \le 0 \\ 100 & \text{if } Y_i^{(\mathrm{math}),*}(z) \ge 100 \\ Y_i^{(\mathrm{math}),*}(z) & \text{otherwise,} \end{cases}$$

where

$$Y_i^{(\mathrm{math}),*}(z) \mid W_i, X_i^{\mathrm{obs}}, Rx_i, C_i, \theta^{(\mathrm{math})} \sim \mathrm{N}\big(g_2(W_i, X_i^{\mathrm{obs}}, Rx_i, C_i, z)'\beta^{(\mathrm{math})}, \exp\big[g_3(X_i^{\mathrm{obs}}, Rx_i, C_i, z)'\zeta^{(\mathrm{math})}\big]\big)$$

for $z = 0, 1$ and $\theta^{(\mathrm{math})} = (\beta^{(\mathrm{math})}, \zeta^{(\mathrm{math})})$. Here $Y_i^{(\mathrm{math}),*}(0)$ and $Y_i^{(\mathrm{math}),*}(1)$ are assumed conditionally independent, an assumption that has no effect on inference for our superpopulation parameters of interest (Rubin 1978a).

The results reported in Section 8 use an outcome component model whose outcome mean link function, g_2, is linear in distinct parameters for the following:

1. For the students of the PMPD design: an intercept, applicant's school (low/high), ethnicity, indicators for grade, an indicator for whether or not the pretest scores were available, pretest score values for the subjects with available scores, and propensity score
2. For the students of the other periods: the variables in item 1, along with indicators for application period
3. An indicator for whether or not a person is an always taker
4. An indicator for whether or not a person is a complier
5. For compliers assigned treatment: an intercept, applicant's school (low/high), ethnicity, and indicators for the first three grades (with the variable for the fourth-grade treatment effect as a function of the already-included variables)

For the variance of the outcome component, the link function g_3 includes indicators that saturate the cross-classification of whether or not a person applied in the PMPD period and whether or not the pretest scores were available. This dependence is needed because each pattern conditions on a different set of covariates; that is, X^{obs} varies from pattern to pattern.

The prior distributions for the outcome submodel are

$$\beta^{(\mathrm{math})} \mid \zeta^{(\mathrm{math})} \sim \mathrm{N}\big(0, F(\zeta^{(\mathrm{math})})\xi\mathbf{I}\big),$$

$$\text{where } F\big(\zeta^{(\mathrm{math})}\big) = \frac{1}{K}\sum_k \exp\big(\zeta_k^{(\mathrm{math})}\big),$$

$\zeta^{(\mathrm{math})} = (\zeta_1^{(\mathrm{math})}, \ldots, \zeta_K^{(\mathrm{math})})$, with one component for each of the K (in our case, $K = 4$) subgroups defined by cross-classifying the PMPD/non-PMPD classification and the missing-data indicator for pretest scores, and where ξ is a hyperparameter set at 10; and $\exp(\zeta_k^{(\mathrm{math})}) \overset{\mathrm{iid}}{\sim} \mathrm{inv}\chi^2(\nu, \sigma^2)$, where $\mathrm{inv}\chi^2(\nu, \sigma^2)$ refers to the distribution of the inverse of a chi-squared random variable with degrees of freedom ν (set at 3) and scale parameter σ^2 (set at 400 based on preliminary estimates of variances). The sets of values for these and the hyperparameters of the other model components were chosen for satisfying two criteria: giving posterior standard deviations for the estimands in the neighborhood of the respective standard errors of approximate maximum likelihood estimate fits of similar preliminary likelihood models and giving satisfactory model checks (Sec. 9.2), and producing quick enough mixing of the Bayesian analysis algorithm.

We specify the marginal distribution for reading outcome $Y^{(\mathrm{read})}$ in the same way as for the math outcome, with separate mean and variance regression parameters $\beta^{(\mathrm{read})}$ and $\zeta^{(\mathrm{read})}$. Finally, we allow that, conditionally on $W_i, X_i^{\mathrm{obs}}, Rx_i, C_i$, the math and reading outcomes at a given assignment, $Y_i^{(\mathrm{math}),*}(z)$ and $Y_i^{(\mathrm{read}),*}(z)$, may be dependent with an unknown correlation, ρ. We set the prior distribution for ρ to be uniform in $(-1, 1)$ independently of the remaining parameters in their prior distributions.

7.3 Outcome Response Submodel

As with the pretests, the outcomes on mathematics and reading were either jointly observed or jointly missing, thus one indicator $Ry_i(z)$ for missingness of outcomes is sufficient for each assignment $z = 0, 1$. For the submodel on this indicator, we also use a probit specification,

$$Ry_i(z) = 1 \quad \text{if } Ry_i(z)^* \equiv g_2(W_i, X_i^{\mathrm{obs}}, Rx_i, C_i, z)'\beta^{(\mathrm{R})} + E_i(z) \ge 0,$$

where $Ry_i(0)$ and $Ry_i(1)$ are assumed conditionally independent (using the same justification as for the potential outcomes) and where $E_i(z) \sim \mathrm{N}(0, 1)$. The link function of the probit model on the outcome response g_2 is the same as the link function for the mean of the outcome component. The prior distribution for the outcome response submodel is $\beta^{(\mathrm{R})} \sim \mathrm{N}(0, \{\sigma^{(\mathrm{R})}\}^2\mathbf{I})$, where $\{\sigma^{(\mathrm{R})}\}^2$ is a hyperparameter set at 10.

8. RESULTS

All results herein were obtained from the same Bayesian analysis. We report results for the ITT and CACE estimands, for proportions of compliance principal strata, and for outcome response rates. The reported estimands are not parameters of the model, but rather are functions of parameters and data. The results are reported by applicant's school (low/high) and grade. Both of these variables represent characteristics of children that potentially could be targeted differentially by government policies. Moreover, each was thought to have possible interaction effects with treatment assignment. Except when otherwise stated, the plain numbers are posterior means, and the numbers in parentheses are 2.5 and 97.5 percentiles of the posterior distribution.

8.1 Test Score Results

Here we address two questions:

1. What is the impact of being offered a scholarship on student outcomes, namely, the ITT estimand?
2. What is the impact of attending a private school on student outcomes, namely, the CACE estimand?

The math and reading posttest score outcomes represent the national percentile rankings within grade. They have been adjusted to correct for the fact that some children were kept behind while others skipped a grade, because students transferring to private schools are hypothesized to be more likely to have been kept behind by those schools. The individual-level causal estimates have also been weighted so that the subgroup causal estimates correspond to the effects for all eligible children belonging to that subgroup who attended a screening session.

8.1.1 Effect of Offering the Scholarship on Mathematics and Reading. We examine the impact of being offered a scholarship (the ITT effect) on posttest scores. The corresponding estimand for individual i is defined as $\mathrm{E}(Y_i(1) - Y_i(0) \mid W_i^{\mathrm{p}}, \theta)$, where W_i^{p} denotes the policy variables grade level and applicant's school (low/high) for that individual. The simulated posterior distribution of the ITT effect is summarized in Table 4. To draw from this distribution, we take the following steps:

Table 4. ITT Effect of Winning the Lottery on Math and Reading Test Scores 1

Grade at application	Applicant's school: Low		Applicant's school: High	
	Reading	Math	Reading	Math
1	$2.3_{(-1.3,5.8)}$	$5.2_{(2.0,8.3)}$	$1.4_{(-4.8,7.2)}$	$5.1_{(.1,10.3)}$
2	$.5_{(-2.6,3.5)}$	$1.3_{(-1.7,4.3)}$	$-.6_{(-6.2,4.9)}$	$1.0_{(-4.3,6.1)}$
3	$.7_{(-2.7,4.0)}$	$3.3_{(-.5,7.0)}$	$-.5_{(-6.0,5.0)}$	$2.5_{(-3.2,8.0)}$
4	$3.0_{(-1.1,7.4)}$	$3.1_{(-1.2,7.2)}$	$1.8_{(-4.1,7.6)}$	$2.3_{(-3.3,7.8)}$
Overall	$1.5_{(-.6,3.6)}$	$3.2_{(1.0,5.4)}$	$.4_{(-4.6,5.2)}$	$2.8_{(-1.8,7.2)}$

NOTE: Year Postrandomization Plain numbers are means, and numbers in parentheses are central 95% intervals of the posterior distribution of the effects on percentile rank.

1. Draw θ and $\{C_i\}$ from the posterior distribution (see App. A).
2. Calculate the expected causal effect $\mathrm{E}(Y_i(1) - Y_i(0) \mid W_i^{\mathrm{p}}, X_i^{\mathrm{obs}}, C_i, Rx_i, \theta)$ for each subject based on the model in Section 7.2.
3. Average the values of step 2 over the subjects in the subclasses of W^{p} with weights reflecting the known sampling weights of the study population for the target population.

These results indicate posterior distributions with mass primarily ($>97.5\%$) to the right of 0 for the treatment effect on math scores overall for children from low applicant schools, and also for the subgroup of first graders. Each effect indicates an average gain of greater than three percentile points for children who won a scholarship. All of the remaining intervals cover 0. As a more general pattern, estimates of effects are larger for mathematics than for reading and larger for children from low-applicant schools than for children from high-applicant schools.

These results for the ITT estimand using the method of this article can also be contrasted with the results using a simpler method reported in Table 5. The method used to obtain these results is the same as that used in the initial MPR study (Peterson, Myers, Howell, and Mayer 1999) but limited to the subset of single-child families and separated out by applicant's school (low/high). This method is based on a linear regression of the posttest scores on the indicator of treatment assignment, ignoring compliance data. In addition, the analysis includes the design variables and the pretest scores and is limited to individuals for whom the pretest and posttest scores are known. Separate analyses are run for math and reading posttest scores, and the missingness of such scores is adjusted by inverse probability weighting. Finally, weights are used to make the results for the study participants generalizable to the population of all eligible single-child families who were screened. (For more details of this method, see the appendix of Peterson et al. 1999.)

Table 5. ITT Effect of Winning the Lottery on Test Scores, Estimated Using the Original MPR Method

Grade at application	Applicant's school: Low		Applicant's school: High	
	Reading	Math	Reading	Math
1	$-1.0_{(-7.1,5.2)}$	$2.1_{(-2.6,6.7)}$	$4.8_{(-10.0,19.6)}$	$2.6_{(-15.5,20.7)}$
2	$-.8_{(-4.9,3.2)}$	$2.0_{(-4.0,8.0)}$	$-3.4_{(-16.5,9.7)}$	$2.7_{(-10.3,15.7)}$
3	$3.2_{(-1.7,8.1)}$	$5.0_{(-.8,10.7)}$	$-8.0_{(-25.4,9.3)}$	$4.0_{(-17.7,25.6)}$
4	$2.7_{(-3.5,8.8)}$	$.3_{(-7.3,7.9)}$	$27.9_{(8.0,47.8)}$	$22.7_{(-1.5,46.8)}$
Overall	$.6_{(-2.0,3.2)}$	$2.0_{(-.8,4.8)}$	$1.1_{(-7.0,9.1)}$	$.3_{(-9.6,10.1)}$

NOTE: Plain numbers are point estimates, and parentheses are 95% confidence intervals for the mean effects on percentile rank.

Table 6. Effect of Private School Attendance on Test Scores

Grade at application	Applicant's school: Low		Applicant's school: High	
	Reading	Math	Reading	Math
1	$3.4_{(-2.0, 8.7)}$	$7.7_{(3.0, 12.4)}$	$1.9_{(-7.3, 10.3)}$	$7.4_{(.2, 14.6)}$
2	$.7_{(-3.7, 5.0)}$	$1.9_{(-2.4, 6.2)}$	$-.9_{(-9.4, 7.3)}$	$1.5_{(-6.2, 9.3)}$
3	$1.0_{(-4.1, 6.1)}$	$5.0_{(-.8, 10.7)}$	$-.8_{(-9.5, 7.7)}$	$4.0_{(-4.9, 12.5)}$
4	$4.2_{(-1.5, 10.1)}$	$4.3_{(-1.6, 10.1)}$	$2.7_{(-6.3, 11.3)}$	$3.5_{(-4.7, 11.9)}$
Overall	$2.2_{(-.9, 5.3)}$	$4.7_{(1.4, 7.9)}$	$.6_{(-7.1, 7.7)}$	$4.2_{(-2.6, 10.9)}$

NOTE: Plain numbers are means, and parentheses are central 95% intervals of the posterior distribution of the effects on percentile rank.

The results of the new method (Table 4) are generally more stable than those of the original method (Table 5), which in some cases are not even credible. In the most extreme cases, the original method estimates a 22.7-point gain [95% confidence interval, $(-1.5, 46.8)$] in mathematics and a 27.9-point gain [95% confidence interval, $(8.0, 47.8)$] in reading for fourth-grade children from high-applicant schools. More generally, the new method's results display a more plausible pattern in comparing effects in high-applicant versus low-applicant schools and mathematics versus reading.

8.1.2 Effect of Private School Attendance on Mathematics and Reading. We also examine the effect of offering the scholarship when focusing only on the compliers (the CACE). The corresponding estimand for individual i is defined as

$$\mathrm{E}(Y_i(1) - Y_i(0) \mid W_i^{\mathrm{p}}, C_i = c, \theta).$$

This analysis defines the treatment as private school attendance (Sec. 6.3). The simulated posterior distribution of the CACE is summarized in Table 6. A draw from this distribution is obtained using steps 1–3 described in the Section 8.1.1 for the ITT estimand, with the exception that at step 3 the averaging is restricted to the subjects whose current draw of C_i is "complier."

The effects of private school attendance follow a pattern similar to that of the ITT effects, but the posterior means are slightly bigger in absolute value than ITT. The intervals have also grown, reflecting that these effects are for only subgroups of all children, the "compliers," in each cell. As a result, the associated uncertainty for some of these effects (e.g., for fourth-graders applying from high-applicant schools) is large.

8.2 Compliance Principal Strata and Missing Outcomes

Table 7 summarizes the posterior distribution of the estimands of the probability of being in stratum C as a function of an applicant's school and grade, $p(C_i = t \mid W_i^{\mathrm{p}}, \theta)$. To draw from this distribution, we use step 1 described in Section 8.1.1 and then calculate $p(C_i = t \mid W_i^{\mathrm{p}}, X_i^{\mathrm{obs}}, Rx_i, \theta)$ for each subject based on the model of Section 7.1 and average these values as in step 3 of Section 8.1.1. The clearest pattern revealed by Table 7 is that for three out of four grades, children who applied from low-applicant schools are more likely to be compliers or always takers than are children who applied from high-applicant schools.

As stated earlier, theoretically under our structural assumptions, standard ITT analyses or standard IV analyses that use ad hoc approaches to missing data are generally not appropriate for estimating the ITT or CACE estimands when the compliance principal strata have differential response (i.e., outcome missing-data) behaviors. To evaluate this here, we simulated the posterior distributions of

$$p(R_i(z) \mid C_i = t, W_i^{\mathrm{p}}, \theta), \qquad z = 0, 1. \tag{1}$$

To draw from the distributions of the estimands in (1), we used step 1 from Section 8.1.1 and then, for $z = 0, 1$ for each subject calculated $p(R_i(z) \mid W_i^{\mathrm{p}}, X_i^{\mathrm{obs}}, Rx_i, C_i, \theta)$. We then averaged these values over subjects within subclasses defined by the different combinations of W_i^{p} and C_i.

For each draw (across individuals) from the distribution in (1) we calculate the odds ratios (a) between compliers attending public schools and never takers, (b) between compliers attending private schools and always takers, and (c) between compliers attending private schools and compliers attending public schools. The results (omitted) showed that response was increasing in the following order: never takers, compliers attending public schools, compliers attending private schools, and always takers. These results confirm that the compliance principal stratum is an important factor in response both alone and in interaction with assignment.

9. MODEL BUILDING AND CHECKING

This model was built through a process of fitting and checking a succession of models. Of note in this model are the following features: a censored normal model that accommodates the

Table 7. Proportions of Compliance Principal Strata Across Grade and Applicant's School

Grade at application	Applicant's school: Low			Applicant's school: High		
	Never taker	Compiler	Always taker	Never taker	Compiler	Always taker
1	$24.5_{(2.9)}$	$67.1_{(3.8)}$	$8.4_{(2.4)}$	$25.0_{(5.0)}$	$69.3_{(6.1)}$	$5.7_{(3.3)}$
2	$20.5_{(2.7)}$	$69.4_{(3.7)}$	$10.1_{(2.5)}$	$25.3_{(5.1)}$	$67.2_{(6.2)}$	$7.5_{(3.4)}$
3	$24.5_{(3.2)}$	$65.9_{(4.0)}$	$9.6_{(2.5)}$	$28.8_{(5.6)}$	$64.1_{(6.7)}$	$7.1_{(3.5)}$
4	$18.4_{(3.3)}$	$72.8_{(4.6)}$	$8.8_{(3.0)}$	$27.0_{(5.5)}$	$66.7_{(6.7)}$	$6.3_{(3.6)}$

NOTE: Plain numbers are means, and parentheses are standard deviations of the posterior distribution of the estimands.

pile-up of 0s in the outcomes, incorporation of heteroscedasticity, and a multivariate model to accommodate both outcomes simultaneously. It is reassuring from the perspective of model robustness that the results from the last few models fit, including those in the conference proceedings report of Barnard et al. (2002), are consistent with the results from this final model.

9.1 Convergence Checks

Because posterior distributions were simulated from a Markov chain Monte Carlo algorithm (App. A), it is important to assess its convergence. To do this, we ran three chains from different starting values. To initiate each chain, we set any unknown compliance stratum equal to a Bernoulli draw, with probabilities obtained from moment estimates of the probabilities of being a complier given observed attendance and randomization data (D, Z) alone. Using the initialized compliance strata for each chain, parameters were initialized to values based on generalized linear model estimates of the model components.

Each chain was run for 15,000 iterations. At 5,000 iterations, and based on the three chains for each model, we calculated the potential scale-reduction statistic (Gelman and Rubin 1992) for the 250 estimands (parameters and functions of parameters) that serve as building blocks for all other estimands. The results suggested good mixing of the chains (with the maximum potential scale reduction statistic across parameters 1.04) and provided no evidence against convergence. Inference is based on the remaining 30,000 iterations, combining the three chains.

9.2 Model Checks

We evaluate the influence of the model presented in Section 7 with six posterior predictive checks, three checks for each of the two outcomes. A posterior predictive check generally involves (a) choosing a discrepancy measure, that is, a function of observed data and possibly of missing data and the parameter vector θ; and (b) computing a posterior predictive p value (PPPV), which is the probability over the posterior predictive distribution of the missing data and θ that the discrepancy measure in a new study drawn with the same θ as in our study would be as or more extreme than in our study (Rubin 1984; Meng 1996; Gelman, Meng, and Stern 1996).

Posterior predictive checks in general, and PPPVs in particular, demonstrate whether the model can adequately preserve features of the data reflected in the discrepancy measure, where the model here includes the prior distribution as well as the likelihood (Meng 1996). As a result, properties of PPPVs are not exactly the same as properties of classical p values under frequency evaluations conditional on the unknown θ, just as they are not exactly the same for frequency evaluations over both levels of uncertainty, that is, the drawing of θ from the prior distribution and the drawing of data from the likelihood given θ. For example, over frequency evaluations of the latter type, a PPPV is stochastically less variable than but has the same mean as the uniform distribution and so tends to be more conservative than a classical p value, although the reverse can be true over frequency evaluations of the first type (Meng 1996). (For more details on the interpretation and properties of the PPPVs, see also Rubin 1984 and Gelman et al. 1996.)

Table 8. Posterior Predictive Checks: p Values

	Signal	*Noise*	*Signal to Noise*
Math	.34	.79	.32
Read	.40	.88	.39

The posterior predictive discrepancy measures that we choose here are functions of

$$A^{\text{rep}}_{p,z} = \left\{ Y^{\text{rep}}_{ip} : \text{I}(Ry^{\text{rep}}_{ip} = 1)\text{I}(Y^{\text{rep}}_{ip} \neq 0)\text{I}(C^{\text{rep}}_{i} = c)\text{I}(Z_i = z) = 1 \right\}$$

for the measures that are functions of data Y^{rep}_{ip}, Ry^{rep}_{ip}, and C^{rep}_{i} from a replicated study and

$$A^{\text{study}}_{p,z} = \left\{ Y_{ip} : \text{I}(Ry_{ip} = 1)\text{I}(Y_{ip} \neq 0)\text{I}(C_i = c)\text{I}(Z_i = z) = 1 \right\}$$

for the measures that are functions of this study's data. Here Ry^{rep}_{ip} is defined so that it equals 1 if Y^{rep}_{ip} is observed and 0 otherwise and p equals 1 for math outcomes and 2 for reading outcomes. The discrepancy measures, "rep" and "study", that we used for each outcome ($p = 1, 2$) were (a) the absolute value of the difference between the mean of $A_{p,1}$ and the mean of $A_{p,0}$ ("signal"), (b) the standard error based on a simple two-sample comparison for this difference ("noise"), and (c) the ratio of (a) to (b) ("signal to noise"). Although these measures are not treatment effects, we chose them here to assess whether the model can preserve broad features of signal, noise, and signal-to-noise ratio in the continuous part of the compliers' outcome distributions, which we think can be very influential in estimating the treatment effects of Section 8. More preferable measures might have been the posterior mean and standard deviation for the actual estimands in Section 8 for each replicated dataset, but this would have required a prohibitive amount of computer memory because of the nested structure of that algorithm. In settings such as these, additional future work on choices of discrepancy measures is of interest.

PPPVs for the discrepancy measures that we chose were calculated as the percentage of draws in which the replicated discrepancy measures exceeded the value of the study's discrepancy measure. Extreme values (close to 0 or 1) of a PPPV would indicate a failure of the prior distribution and likelihood to replicate the corresponding measure of location, dispersion, or their relative magnitude and would indicate an undesirable influence of the model in estimation of our estimands. Results from these checks, displayed in Table 8, provide no special evidence for such influences of the model.

10. DISCUSSION

In this article we have defined the framework for principal stratification in broken randomized experiments to accommodate noncompliance, missing covariate information, missing outcome data, and multivariate outcomes. We make explicit a set of structural assumptions that can identify the causal effects of interest, and we also provide a parametric model that is appropriate for practical implementation of the framework in settings such as ours.

Results from our model in the school choice study do not indicate strong treatment effects for most of the subgroups examined. But we do estimate positive effects (on the order of 3 percentile points for ITT and 5 percentile points for the effect of

attendance) on math scores overall for children who applied to the program from low-applicant schools, particularly for first graders. Also, the effects were larger for African-American children than for the remaining children (App. B).

Posterior distributions for the CACE estimand, which measures the effect of attendance in private school versus public school, are generally larger than the corresponding ITT effects but are also associated with greater uncertainty. Of importance, because of the missing outcomes, a model like ours is needed for valid estimation even of the ITT effect.

The results from this randomized study are not subject to selection bias in the way that nonrandomized studies in school choice have been. Nevertheless, although we use the CACE, a well-defined causal effect, to represent the effect of attendance of private versus public schools, it is important to remember that the CACE is defined on a subset of the study children (those who would have complied with either assignment) and that for the other children there is no information on such an effect of attendance in this study. Therefore, as with any randomized trial based on a subpopulation, external information, such as background variables, also must be used when generalizing the CACE from compliers to other target groups of children. Also, it is possible that a broader implementation of a voucher program can have a collective effect on the public schools if a large shift of the children who might use vouchers were to have an impact on the quality of learning for children who would stay in public schools (a violation of the no-interference assumption). Because our study contains a small fraction of participants relative to all school children, it cannot provide direct information about any such collective effect, and additional external judgment would need to be used to address this issue.

The larger effects that we estimated for children applying from schools with low versus high past scores are also not, in principle, subject to the usual regression to the mean bias, in contrast to a simple before–after comparison. This is because in our study both types of children are randomized to be offered the scholarship or not, and both types in both treatment arms are evaluated at the same time after randomization. Instead, the differential effect for children from schools with different past scores is evidence supporting the claim that the school voucher program has greater potential benefit for children in lower-scoring schools.

Our results also reveal differences in compliance and missing-data pattern across groups. Differences in compliance indicate that children applying from low-applicant schools are generally more likely to comply with their treatment assignment; this could provide incentives for policy makers to target this subgroup of children. However, this group also exhibits higher levels of always takers, indicating that parents of children attending these poorly performancing schools are more likely to get their child into a private school even in the absence of scholarship availability. These considerations would of course have to be balanced with (and indeed might be dwarfed by) concerns about equity. Missing-data patterns reveal that perhaps relatively greater effort is needed in future interventions of this nature to retain study participants who stay in public schools, particularly the public-school students who actually won but did not use a scholarship.

The approach presented here also has some limitations. First, we have presented principal stratification only on a binary controlled factor Z and a binary uncontrolled factor D, and it is of interest to develop principal stratification for more levels of such factors. Approaches that extend our framework in that direction, including for time-dependent data, have been proposed (e.g., Frangakis et al. 2002b).

Also, the approach of explicitly conditioning on the patterns of missing covariates that we adopted in the parametric component of Section 7 is not as applicable when there are many patterns of missing covariates. In such cases, it would be more appropriate to use models related to those of D'Agostino and Rubin (2000) and integrate them with principal stratification for noncompliance. Moreover, it would be also interesting to investigate results for models fit to these data that allow deviations from the structural assumptions, such as weakened exclusion restrictions (e.g., Hirano et al. 2000; Frangakis, Rubin, and Zhou 2002a), although to ensure robustness of such models, it would be important to first investigate and rely on additional alternative assumptions that would be plausible.

APPENDIX A: COMPUTATIONS

Details are available at *http://biosun01.biostat.jhsph.edu/~cfrangak/papers/sc.*

Table B.1. ITT Estimand

Grade at application	Ethnicity	Applicant school: Low		Applicant school: High	
		Reading	Math	Reading	Math
1	AA	$2.8_{(-1.5,7.2)}$	$6.7_{(3.0,10.4)}$	$1.8_{(-5.2,8.4)}$	$6.3_{(.8,11.9)}$
	Other	$1.7_{(-1.9,5.4)}$	$3.9_{(.5,7.2)}$	$.8_{(-4.5,6.1)}$	$3.5_{(-1.2,8.2)}$
2	AA	$.8_{(-2.8,4.4)}$	$2.4_{(-1.0,5.8)}$	$-.3_{(-6.6,6.0)}$	$2.2_{(-3.8,8.0)}$
	Other	$.2_{(-3.1,3.4)}$	$.5_{(-2.8,3.7)}$	$-.9_{(-6.1,4.3)}$	$0_{(-5.2,4.8)}$
3	AA	$1.1_{(-3.0,5.2)}$	$4.6_{(.3,8.9)}$	$-.1_{(-6.7,6.4)}$	$4.2_{(-2.5,10.6)}$
	Other	$.3_{(-3.2,3.8)}$	$2.0_{(-2.0,5.8)}$	$-.7_{(-5.8,4.4)}$	$1.4_{(-3.9,6.5)}$
4	AA	$3.7_{(-1.2,8.7)}$	$4.4_{(-.5,9.3)}$	$2.4_{(-4.7,9.4)}$	$3.8_{(-2.9,10.5)}$
	Other	$2.4_{(-1.7,6.7)}$	$1.8_{(-2.6,6.0)}$	$1.4_{(-4.3,7.0)}$	$1.1_{(-4.4,6.4)}$
Overall	AA	$2.0_{(-.9,4.8)}$	$4.5_{(1.8,7.2)}$	$.8_{(-5.1,6.5)}$	$4.2_{(-1.1,9.3)}$
	Other	$1.1_{(-1.4,3.6)}$	$2.1_{(-.7,4.7)}$	$.1_{(-4.6,4.6)}$	$1.6_{(-2.9,5.8)}$

NOTE: Plain numbers are means, and parentheses are central 95% intervals of the posterior distribution of the effects on percentile rank.

Table B.2. CACE Estimand

Grade at application	Ethnicity	Applicant school: Low		Applicant school: High	
		Reading	Math	Reading	Math
1	AA	$3.8_{(-2.0,9.6)}$	$9.0_{(4.1,14.0)}$	$2.3_{(-7.3,10.9)}$	$8.3_{(1.1,15.5)}$
	Other	$2.9_{(-3.1,8.8)}$	$6.3_{(.8,11.8)}$	$1.3_{(-8.0,10.2)}$	$5.9_{(-2.2,13.9)}$
2	AA	$1.1_{(-3.6,5.7)}$	$3.1_{(-1.3,7.5)}$	$-.4_{(-9.0,7.9)}$	$2.9_{(-5.0,10.7)}$
	Other	$.3_{(-4.9,5.4)}$	$.8_{(-4.4,5.9)}$	$-1.5_{(-10.5,7.3)}$	$0_{(-8.5,8.5)}$
3	AA	$1.5_{(-4.1,7.0)}$	$6.3_{(.3,12.2)}$	$-.2_{(-9.1,8.5)}$	$5.6_{(-3.2,14.1)}$
	Other	$.5_{(-5.4,6.4)}$	$3.5_{(-3.4,10.2)}$	$-1.3_{(-10.5,7.9)}$	$2.7_{(-7.0,12.0)}$
4	AA	$4.6_{(-1.6,10.8)}$	$5.5_{(-.7,11.6)}$	$3.0_{(-6.0,11.6)}$	$4.8_{(-3.6,13.1)}$
	Other	$3.7_{(-2.6,10.1)}$	$2.8_{(-3.8,9.3)}$	$2.3_{(-7.4,11.7)}$	$2.0_{(-7.0,11.0)}$
Overall	AA	$2.6_{(-1.2,6.3)}$	$6.0_{(2.4,9.5)}$	$1.0_{(-6.9,8.4)}$	$5.5_{(-1.5,12.2)}$
	Other	$1.7_{(-2.3,5.8)}$	$3.3_{(-1.2,7.7)}$	$.1_{(-8.1,7.9)}$	$2.7_{(-5.0,10.3)}$

NOTE: Plain numbers are means, and parentheses are central 95% intervals of the posterior distribution of the effects on percentile rank.

APPENDIX B: ETHNIC BREAKDOWN

The model of Section 7 can be used to estimate the effect of the program on finer strata that may be of interest. For example, to estimate effects stratified by ethnicity (AA for African-American), we obtain the posterior distribution of the causal effect of interest (ITT or CACE) using the same steps described in Section 8.1.1 or 8.1.2, but allowing ethnicity to be part of the definition of W^{P}. The results for this stratification are reported in Table B.1 for the estimands of ITT and Table B.2 for the estimands of CACE.

The results follow similar patterns to those in Section 8. For each grade and applicant's school (low/high) combination, however, the effects are more positive on average for the subgroup of African-American children. For both estimands, this leads to 95% intervals that are entirely above 0 for math scores for the following subgroups: African-American first- and third-graders and overall from low-applicant schools, African-American first-graders from high-applicant schools, and non-African-American first-graders from low-applicant schools. All intervals for reading covered the null value. This suggests that the positive effects reported on math scores in Section 8 for children originating from low-applicant schools are primarily attributable to gains among the African-American children in this subgroup.

[Received October 2002. Revised October 2002.]

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–472.

Barnard, J., Du, J., Hill, J. L., and Rubin, D. B. (1998), "A Broader Template for Analyzing Broken Randomized Experiments," *Sociological Methods and Research*, 27, 285–317.

Barnard, J., Frangakis, C., Hill, J. L., and Rubin, D. B. (2002), "School Choice in NY City: A Bayesian Analysis of an Imperfect Randomized Experiment," in *Case Studies in Bayesian Statistics*, vol. V, eds. Gatsonis, C., Kass, R. E., Carlin, B., Cariiguiry, A., Gelman, A., Verdinelli, I., West, M. New York: Springer-Verlag, pp. 3–97.

Brandl, J. E. (1998), *Money and Good Intentions Are Not Enough, or Why Liberal Democrats Think States Need Both Competition and Community*, Washington, DC: Brookings Institute Press.

Carnegie Foundation for the Advancement of Teaching (1992), *School Choice: A Special Report*, San Francisco: Jossey-Bass.

Chubb, J. E., and Moe, T. M. (1990), *Politics, Markets and America's Schools*, Washington, DC: Brookings Institute Press.

Cobb, C. W. (1992), *Responsive Schools, Renewed Communities*, San Francisco: Institute for Contemporary Studies.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1982), *High School Achievement*, New York: Basic Books.

Cookson, P. W. (1994), *School Choice: The Struggle for the Soul of American Education*, New Haven, CT: Yale University Press.

Coulson, A. J. (1999), *Market Education: The Unknown History*, Bowling Green, OH: Social Philosophy & Policy Center.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

D'Agostino, Ralph B., J. and Rubin, D. B. (2000), "Estimating and Using Propensity Scores With Incomplete Data," *Journal of the American Statistical Association*, 95, 749–759.

Education Week (1998), *Quality Counts '98: The Urban Challenge; Public Education in the 50 States*, Bethesda, MD: Editorial Projects in Education.

Frangakis, C. E., and Rubin, D. B. (1999), "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes," *Biometrika*, 86, 365–380.

——— (2001), "Addressing the Idiosyncrasy in Estimating Survival Curves Using Double-Sampling in the Presence of Self-Selected Right Censoring" (with discussion), *Biometrics*, 57, 333–353.

——— (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 20–29.

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002a), "Clustered Encouragement Design With Individual Noncompliance: Bayesian Inference and Application to Advance Directive Forms" (with discussion), *Biostatistics*, 3, 147–164.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Mahboobeh, S., Vlahov, D., and Strathdee, S. A. (2002b), "Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, With Application to a Needle Exchange Program," Technical Report NEP-06-02, Johns Hopkins University, Dept. of Biostatistics.

Fuller, B., and Elmore, R. F. (1996), *Who Chooses? Who Loses? Culture, Institutions, and the Unequal Effects of School Choice*, New York: Teachers College Press.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (Disc: P760-807)," *Statistica Sinica*, 6, 733–760.

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–472.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993), "Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-Ups," *Journal of the American Statistical Association*, 88, 984–993.

Goldberger, A. S., and Cain, G. G. (1982), "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report," *Sociology of Education*, 55, 103–122.

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12.

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hill, J. L., Rubin, D. B., and Thomas, N. (2000), "The Design of the New York School Choice Scholarship Program Evaluation," in *Donald Campbell's Legacy*, ed. L. Bickman, Newbury Park, CA: Sage Publications.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, A. (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88.

Holland, P. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 396, 945–970.

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Imbens, G. W., and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *The Annals of Statistics*, 25, 305–327.
Levin, H. M. (1998), "Educational Vouchers: Effectiveness, Choice, and Costs," *Journal of Policy Analysis and Management*, 17, 373–392.
Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.
——— (1996), "Pattern-Mixture Models for Multivariate Incomplete Data With Covariates," *Biometrics*, 52, 98–111.
Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.
Meng, X. L. (1996), "Posterior Predictive p Values," *The Annals of Statistics*, 22, 1142–1160.
Neal, D. (1997), "The Effects of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics*, 15, 98–123.
Neyman, J. (1923), "On the Application of Probablity Theory to Agricultural Experiments Essay on Principles. Section 9," translated in *Statistical Science*, 5, 465–480.
Peterson, P. E., and Hassel, B. C. (Eds.) (1998), *Learning from School Choice*, Washington, DC: Brookings Institute Press.
Peterson, P. E., Myers, D. E., Howell, W. G., and Mayer, D. P. (1999), "The Effects of School Choice in New York City," in *Earning and Learning; How Schools Matter*, eds. S. E. Mayer and P. E. Peterson, Washington, DC: Brookings Institute Press.
Robins, J. M., Greenland, S., and Hu, F.-C. (1999). "Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome" (with discussion), *Journal of the American Statistical Association*, 94, 687–712.
Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.
Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.
——— (1977), "Assignment to Treatment Groups on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.
——— (1978a), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.
——— (1978b), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse (C/R: P29-34)," in *Proceedings of the Survey Research Methods Section, American Statistical Association*. pp. 20–28.
——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.
——— (1980). Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," *Journal of the American Statistical Association*, 75, 591–593.
——— (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
——— (1990), "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.
The Coronary Drug Project Research Group (1980), "Influence of Adherence to Treatment and Response of Cholesterol on Mortality in the Coronary Drug Project," *New England Journal of Medicine*, 303, 1038–1041.
Tinbergen, J. (1930), "Determination and Interpretation of Supply Curves: An Example," in *The Foundations of Econometric Analysis*, eds. D. Hendry and M. Morgan, Cambridge, U.K: Cambridge University Press.
Wilms, D. J. (1985), "Catholic School Effect on Academic Achievement: New Evidence From the High School and Beyond Follow-Up Study," *Sociology of Education*, 58, 98–114.

Comment

Bengt MUTHÉN, Booil JO, and C. Hendricks BROWN

1. INTRODUCTION

The article by Barnard, Frangakis, Hill, and Rubin (BFHR) is timely in that the Department of Education is calling for more randomized studies in educational program evaluation. (See the discussion of the "No Child Left Behind" initiative, in e.g., Slavin 2002.) BFHR can serve as a valuable pedagogical example of a successful sophisticated statistical analysis of a randomized study. Our commentary is intended to provide additional pedagogical value to benefit the planning and analysis of future studies, drawing on experiences and research within our research group. [The Prevention Science Methodology Group (PSMG; *www.psmg.hsc.usf.edu*), co-PI's Brown and Muthén, has collaborated over the last 15 years with support from the National Institute of Mental Health and the National Institute on Drug Abuse.]

BFHR provides an exemplary analysis of the data from an imperfect randomized trial that suffers from several complications simultaneously: noncompliance, missing data in outcomes, and missing data in covariates. We are very pleased to see their application of cutting-edge Bayesian methods for dealing with these complexities. In addition, we believe the methodological issues and the results of the study have important implications for the design and analysis of randomized trials in education and for related policy decisions.

BFHR provides results of the New York City school choice experiment based on 1-year achievement outcomes. With the planned addition of yearly follow-up data, growth models can provide an enhanced examination of causal impact. We discuss how such growth modeling can be incorporated and provide a caution that applies to BFHR's use of only one posttest occasion. We also consider the sensitivity of the latent class ignorability assumption in combination with the assumption of compound exclusion.

2. LONGITUDINAL MODELING ISSUES

BFHR focuses on variation in treatment effect across compliance classes. This part of the commentary considers variation in treatment effect across a different type of class based on the notion that the private school treatment effect might very well be quite different for children with different achievement development. (Also of interest is potential variation in treatment effects across schools, with respect to both the public school the child originated in and the private school the child was

Bengt Muthén is Professor and Booil Jo is postdoc, Graduate School of Education and Information Studies, University of California Los Angeles, Los Angeles, CA-90095 (E-mail: *bmuthen@ucla.edu*). C. Hendricks Brown is Professor, Department of Epidemiology and Biostatistics, University of South Florida, Tampa, FL 33620. The research of the first author was supported by National Institute on Alcohol Abuse and Alcoholism grant K02 AA 00230. The research of the second and third authors was supported by National Institute on Drug Abuse and National Institute of Mental Health grant MH40859. The authors thank Chen-Pin Wang for research assistance, Joyce Chappell for graphical assistance, and the members of the Prevention Science Methodology Group and the Fall ED299A class for helpful comments.

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Applications and Case Studies
DOI 10.1198/016214503000080

moved to, but this multilevel aspect of the data is left aside here for lack of space.) To study such a "treatment–baseline interaction" (or "treatment–trajectory interaction"), we will switch from BFHR's pretest–posttest analysis framework (essentially a very advanced ANCOVA-type analysis) to the growth mixture modeling framework of Muthén et al. (2002). An underlying rationale for this modeling is that individuals at different initial status levels, and on different trajectories, may benefit differently from a given treatment. ANCOVA controls for initial status, as measured by the observed pretest score. Unlike the observed pretest score, the latent variable of initial status is free of time-specific variation and measurement error.

The focus on longitudinal aspects of the New York School Choice Study (NYSCS) is both substantively and statistically motivated. First, treatment effects may not have gained full strength after only a 1-year stay in a private school. The NYSCS currently has data from three follow-ups, that is, providing repeated-measures data from four grades. Although BFHR used percentile scores that do not lend themselves to growth modeling, a conversion to "scale scores" (i.e., IRT-based, equated scores) should be possible, enabling growth modeling. Unfortunately, educational research traditionally uses scales that are unsuitable for growth modeling, such as percentile scores, normal curve equivalents, and grade equivalents (for a comparison in a growth context, see Seltzer, Frank, and Bryk 1994). Hopefully, this tradition can be changed. Second, the use of information from more time points than pretest and posttest makes it possible to identify and estimate models that give a richer description of the normative development in the control group and how the treatment changes this development.

Consider three types of latent variables for individual i. The first type, C_i, refers to BFHR's compliance principal strata. The next two relate to the achievement development as expressed by a growth mixture model: T_i refers to trajectory class and η_i refers to random effects within trajectory class (within-class model is a regular mixed-effects model). Unlike the latent class variable C_i, the latent class variable T_i is a fully unobserved variable as is common in latent variable modeling (see, e.g., Muthén 2002a). Consider the likelihood expression for individual i, using the [] notation to denote probabilities/densities,

$$[C_i, T_i|X_i][\eta_i|C_i, T_i, X_i][Y_i|\eta_i, C_i, T_i, X_i] \times [U_i|\eta_i, C_i, T_i, X_i][R_i|Y_i, \eta_i, C_i, T_i, X_i], \quad (1)$$

where X_i denotes covariates, U_i denotes a compliance stratum indicator (with C_i perfectly measured by U_i in the treatment group for never-takers and perfectly measured in the control group for always-takers, other group–class combinations having missing data), and R_i denotes indicators for missing data on the repeated-measures outcomes Y_i (pretreatment and posttreatment achievement scores). This type of model can be fitted into the latent variable modeling framework of the Mplus program (Muthén and Muthén 1998–2002; tech. app. 8), which has implemented an EM-based maximum likelihood estimator. (For related references, see the Mplus website *www.statmodel.com.*) As a special case of (1), conventional random-effects growth modeling includes η_i, but excludes C_i and T_i and assumes missingness at random, so that the last term in (1) is ignored. Growth mixture modeling (Muthén and Shedden 1999; Muthén et al. 2002; Muthén and Muthén 1998–2002) includes η_i and T_i. BFHR includes C_i, but not T_i (or η_i), and includes the last term in (1), drawing on latent ignorability of Frangakis and Rubin (1999). Muthén and Brown (2001) studied latent ignorability related to T_i in the last term of (1). In randomized studies, it would be of interest to study C_i and T_i classes jointly, because individuals in different trajectory classes may show different compliance and missingness may be determined by these classes jointly.

If data have been generated by a growth mixture model with treatment effects varying across trajectory classes, what would pretest–posttest analysis such as that in BFHR reveal? To judge the possibility of such treatment–trajectory interaction in the NYSCS, we considered several recent applications of growth mixture modeling that have used T_i to represent qualitatively different types of trajectories for behavior and achievement scores on children in school settings. Drawing on these real-data studies, two growth mixture scenarios were investigated. (A detailed description of these real-data studies and scenarios and their parameter values are given in Mplus Web Note #5 at *www.statmodel.com/mplus/examples/webnote.html.*) For simplicity, no missing data on the outcome or pretest is assumed and C_i classes are not present. In a three-class scenario, the treatment effect is noteworthy only for a 70% middle class, assuming that the low-class membership (10%) hinders individuals from benefiting from the treatment and assuming that the high-class membership (20%) does not really need the treatment. The achievement development in the three-class scenario is shown in Figure 1(a), and the corresponding posttest (y_2)–pretest (y_1) regressions are shown in Figure 1(b). The lines denoted ANCOVA show a regular ANCOVA analysis allowing for an interaction between treatment and pretest (different slopes). In the three-class scenario, the ANCOVA interaction is not significant at $n = 2{,}000$ and the treatment effect in the middle class is underestimated by 20%, but overestimated in the other two classes. In a two-class scenario (not shown here), where the treatment is noteworthy only for individuals in the low class (50%), ANCOVA detects an interaction that is significant at the NYSCS sample size of $n = 2{,}000$, but underestimates the treatment effect for most children in the low class. (At the low-class average pretest value of 0, the treatment effect is underestimated by 32%.)

Although the NYSCS children are selected from low-performing schools (the average NYSCS math and reading percentile rankings are around 23–28), there may still be sufficient heterogeneity among children in their achievement growth to make a treatment-trajectory interaction plausible. The three-class scenario is possible, perhaps with more children in the low class relative to the other two classes. If this is the case, the ANCOVA analysis shown in Figure 1 suggests a possible reason for BFHR's finding of low treatment effects. The empirical studies and the results in Figure 1 suggest that future program evaluations may benefit from exploring variation in treatment effects across children characterized by different development. Using data from at least two posttreatment time points (three time points total), the class-specific treatment effects generated in these data can be well recovered by growth mixture modeling. (Monte Carlo simulation results are given in Mplus Web Note #5 at *www.statmodel.com/mplus/examples/webnote.html.*)

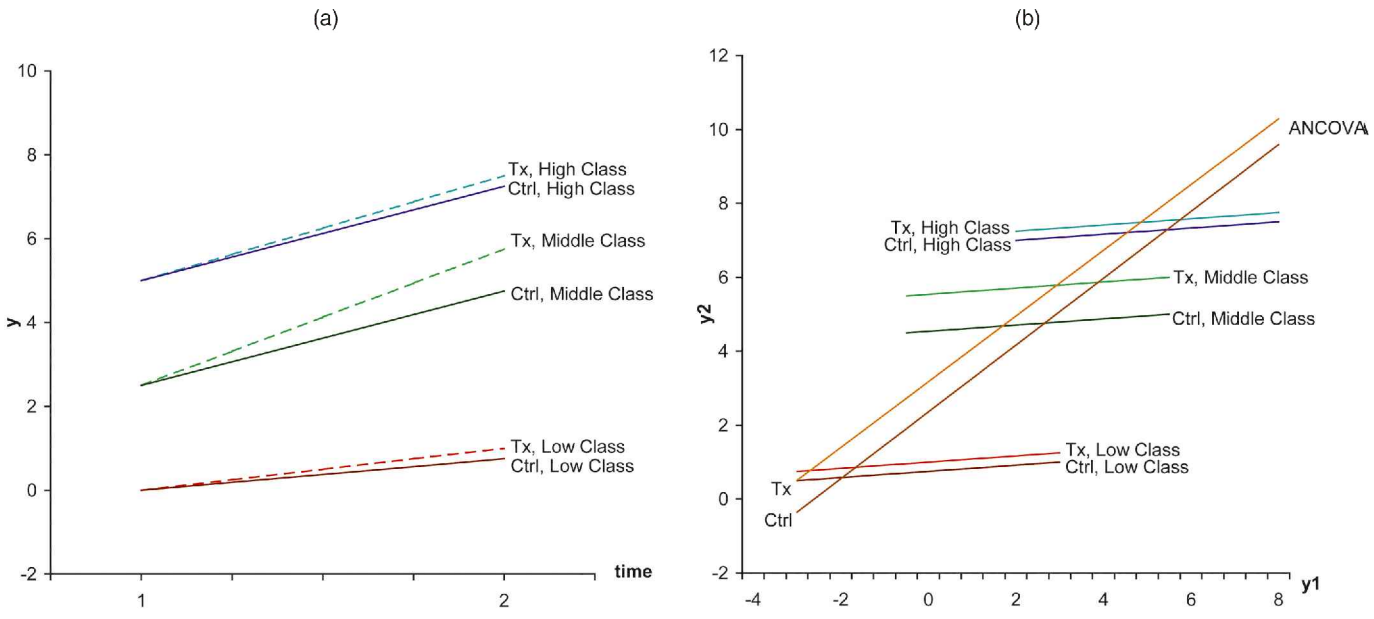

Figure 1. Growth Mixture Modeling Versus Pretest–Posttest Analysis.

A more flexible analysis is obtained with more posttreatment time points. An improved design for the determination of the latent trajectory classes is to use more than one pretreatment time point so that the trajectory class membership is better determined before the treatment starts.

3. COMPOUND EXCLUSION AND LATENT IGNORABILITY

Based on the ideas of principal stratification (Frangakis and Rubin 2002) and latent ignorability (Frangakis and Rubin 1999), BFHR successfully demonstrates that the complexities of educational studies can be better handled under more explicit and flexible sets of assumptions. Although we think that their structural assumptions are reasonable in the NYSCS, we would like to add some thoughts on the plausibility of two other assumptions, considering more general situations.

Compound exclusion (CE) is one of the key structural assumptions in identifying principal effects under latent ignorability. However, the plausibility of this assumption can be questioned in practice (Frangakis et al. 2002; Hirano et al. 2000; Jo, 2002, 2002c; Shadish, Cook, and Campbell, 2002; West and Sagarin, 2000). In the NYSCS, it seems realistic to assume that winning a lottery has some positive impact on always-takers; however, it is less clear how winning a lottery will affect never-takers. One possibility is that winning a lottery has a negative impact on parents, because they fail to benefit from it. Discouraged parents may have a negative influence on a child's test scores or response behaviors. This negative effect may become more evident if noncompliance is due to parents' low expectation of or lack of interest in their child's education. Another possibility is that winning a lottery has a positive impact on a child's test scores or response behaviors. For example, parents who are discouraged by being unable to send their child to private schools even with vouchers may try harder to improve the quality of existing resources (e.g., in the public school their child attends) and be more motivated to support their child to improve his or her academic performance. Given these competing possibilities, it is not easy to predict whether and how CE is violated.

Depending on the situation, causal effect estimates can be quite sensitive to violation of the exclusion restriction in outcome missingness (Jo 2002b), which is less known than the impact of violating exclusion restriction in observed outcomes (Angrist et al. 1996; Jo 2002). The implication of possible violation of CE and its impact is that the relative benefit of models assuming latent ignorability (LI) and standard ignorability (SI) depends on degrees of deviation from CE and SI. Identification of causal effects under LI relies on the generalized (compound) exclusion restriction (i.e., both on the outcomes and missingness of outcomes), whereas identification of causal effects under SI relies on the standard exclusion restriction (i.e., only on the outcomes). Therefore, in some situations, the impact of deviation from CE may outweigh the impact of deviation from SI, resulting in more biased causal effect estimates in models assuming LI than in models assuming SI (Jo 2002b). For example, if SI holds but CE is seriously violated (say, a 20% increase in the response rate due to treatment assignment for compliers and a 15% increase for never-takers), causal effect estimates and the coverage probability assuming LI and CE can drastically deviate from the true value and the nominal level. This type of violation does not affect models assuming SI and the standard exclusion restriction. To empirically examine the plausibility of SI, LI, and CE, it will be useful to do sensitivity analyses of models imposing different combinations of these assumptions. As BFHR points out, this investigation can be conducted by relaxing compound exclusion (e.g., Frangakis et al. 2002; Hirano et al. 2000), or by using alternative structural assumptions (e.g., Jo, 2002c). More research is needed to examine the efficiency of these alternative models and to explore factors associated with insensitivity of LI models to violation of compound exclusion.

4. CONCLUSION

Causal inferences of the type BFHR provides are a dramatic improvement over the existing literature now available on the

question of whether school choice will produce better achievement outcomes for children in an urban public school system. The randomized lottery provides an exceptionally powerful tool for examining the impact of a program—far more useful than observational studies that have causal change intertwined hopelessly with self-selection factors. Statisticians are just now investigating variations in such principal strata analyses, that is, those involving latent classes formed as a function of randomized trials involving intervention invitations (such as vouchers), encouragement designs, and field trial designs involving more than one randomization (Brown and Liao 1999). The latent categories in this article, which BFHR labels "complier," "never-taker," "always-taker," and "defier," represent only one type of design. Other terms may be more relevant to the scientific questions underlying trials in which subjects are randomly assigned to different levels of invitation (e.g., Angrist and Imbens 1995), or different levels of implementation. Such trials not only have great potential for examining questions of effectiveness, sustainability, and scalability, but also require terms more consistent with adherence than compliance. Again, we congratulate the authors on an important addition to the methodological literature that we predict will have lasting impact.

ADDITIONAL REFERENCES

Angrist, J. D., and Imbens, G. W. (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Brown, C. H., and Liao, J. (1999), "Principles for Designing Randomized Preventive Trials in Mental Health: An Emerging Developmental Epidemiologic Perspective," *American Journal of Community Psychology*, 27, 673–709.

Jo, B. (2002a), "Model Misspecification Sensitivity Analysis in Estimating Causal Effects of Interventions With Noncompliance," *Statistics in Medicine*, 21, 3161–3181.

——— (2002b), "Sensitivity of Causal Effects Under Ignorable and Latent Ignorable Missing-Data Mechanisms," available at *www.statmodel.com/mplus/examples/jo/*.

——— (2002c), "Estimation of Intervention Effects With Noncompliance: Alternative Model Specifications" (with discussion), *Journal of Educational and Behavioral Statistics*, 27, 385–420.

Muthén, B. (2002a), "Beyond SEM: General Latent Variable Modeling," *Behaviormetrika*, 29, 81–117.

Muthén, B., and Brown, C. H. (2001), "Non-Ignorable Missing Data in a General Latent Variable Modeling Framework," presented at the annual meeting of the Society for Prevention Research, Washington, DC, June 2001.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S., Carlin, J., and Liao, J. (2002), "General Growth Mixture Modeling for Randomized Preventive Interventions," *Biostatistics*, 3, 459–475.

Muthén, L., and Muthén, B. (1998–2002), *Mplus User's Guide*, Los Angeles: authors.

Muthén, B., and Shedden, K. (1999), "Finite Mixture Modeling With Mixture Outcomes Using the EM Algorithm," *Biometrics*, 55, 463–446.

Seltzer, M. H., Frank, K. A., and Bryk, A. S. (1993), "The Metric Matters: The Sensitivity of Conclusions About Growth in Student Achievement to Choice of Metric," *Educational Evaluation and Policy Analysis*, 16, 41–49.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin.

Slavin, R. E. (2002), "Evidence-Based Education Policies: Transforming Educational Practice and Research," *Educational Researcher*, 31, 15–21.

West, S. G., and Sagarin, B. J. (2000), "Participant Selection and Loss in Randomized Experiments," in *Research Design*, ed. L. Bickman, Thousand Oaks, CA: Sage, pp. 117–154.

Comment

Alan B. KRUEGER and Pei ZHU

The New York City school voucher experiment provides some of the strongest evidence on the effect of private school vouchers on student test achievement yet available. Barnard et al. provide a useful reevaluation of the experiment, and some of the authors were integrally involved in the design of the experiment. We will leave it to other discussants to comment on the Bayesian methodological advances in their paper, and instead comment on the substantive lessons that can be learned from the experiment, and the practical lessons that emerge from the novel design of the experiment.

We have the advantage of having access to more complete data than Barnard et al. used in preparing their paper. This includes more complete baseline test information, data on multichild families as well as single-child families, and three years of follow-up test data instead of just one year of follow-up data. In our comment, we use the more comprehensive sample because results for this sample are more informative, but we highlight where differences arise from using the sample analyzed by Barnard et al.

Three themes emerge from our analysis. First, simplicity and transparency are under appreciated virtues in statistical analysis. Second, it is desirable to use the most recent, comprehensive data, for the widest sample possible. Third, there is no substitute for probing the definitions and concepts that underlie the data.

1. RANDOM ASSIGNMENT

As Barnard et al. explain, the voucher experiment entailed a complicated block design, and different random assignment procedures were used in the first and second set of blocks. In the first block, a propensity matched-pairs design (PMPD) method was used. In this block, far more potential control families were available than was money to follow them up. Rather than select a random sample of controls for follow up after randomly

Alan B. Krueger is Professor and Pei Zhu is graduate student, Economics Department, Princeton University, Princeton, NJ 08544 (E-mail: *akrueger@princeton.edu*). In part, this comment extends and summarizes results of Krueger and Zhu (2003). Readers are invited to read that article for a more in-depth analysis of many of the issues raised in this comment. Some of the results presented here differ slightly from those in our earlier article, however, because the definition of family size in this comment corresponds to the one used by Barnard et al. rather than to the definition in our earlier article.

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Applications and Case Studies
DOI 10.1198/016214503000099

Table 1. Efficiency Comparison of Two Methods of Random Assignment Estimated Treatment Effect and Standard Error, by Method of Random Assignment

		Propensity score match subsample		Randomized blocks subsample		Relative sampling variance	Sample-size-adjusted relative sampling variance
Year	Model	Coefficient	Number of observations	Coefficient	Number of observations	(RB/PMPD)	(RB/PMPD)
All students							
Year 1	Control for baseline	.33 (1.40)	721	2.10 (1.38)	734	.972	.989
	Omit baseline	.32 (1.40)	1,048	−1.02 (1.58)	1,032	1.274	1.254
Year 3	Control for baseline	1.02 (1.56)	613	.67 (1.74)	637	1.244	1.293
	Omit baseline	−.31 (1.57)	900	−.45 (1.78)	901	1.285	1.287
African-American students							
Year 1	Control for baseline	.32 (1.88)	302	8.10 (1.71)	321	.827	.879
	Omit baseline	2.10 (1.99)	436	3.19 (2.07)	447	1.082	1.109
Year 3	Control for baseline	4.21 (2.51)	247	6.57 (2.09)	272	.693	.764
	Omit baseline	2.26 (2.47)	347	3.05 (2.31)	386	.875	.973

NOTE: Standard errors are in parentheses. Treatment effect coefficient is from a regression of test scores on a dummy indicating assignment to receive a voucher (1 = yes), lottery randomization strata dummies, and in indicated models baseline test scores. Bootstrap standard errors account for within-family correlation in residuals. Year 1 or 3 refers to follow-up year. The last column adjusts the relative sampling variances of the treatment effects for differences in relative sample sizes. The sample of African-American students consists of those whose mothers' race/ethnicity is identified as non-Hispanic, Black/African-American.

selecting the treatments, the researchers estimated a propensity score model to align controls with the treatments, and then used a nearest-available-neighbor Mahalanobis match to select specific paired control families for follow-up. In principle, the PMPD design should improve the precision of the estimates by reducing the chance imbalances that can occur in a simple randomized block design. A standard randomized block design was used in the remaining four blocks.

Barnard et al. emphasize that the PMPD block was more balanced for 15 of the 21 baseline variables that they examined. But the key question is the extent to which the precision of the estimates was improved, not the covariate balance, because both blocks yield unbiased estimates. If the covariates explain relatively little of the outcome variable, then the PMPD will add very little. In Table 1 we provide standard intent-to-treat (ITT) estimates—that is, the difference in mean test scores between treatments and controls, conditional on the strata used for random assignment (family size by block by high/low-achieving school)—and, more importantly, their standard errors, separately for the PMPD subsample and randomized block subsample. The outcome variable is the average of the national percentile rank on the math and reading segments of the Iowa Test for Basic Skills, taken either 1 year or 3 years after random assignment. Two sets of results are presented, one set controlling for baseline test scores for the subsample with baseline scores, and the other without controlling for scores for the larger sample. The last column adjusts the relative sampling variances for differences in sample size between the blocks.

For all students (panel A), the standard errors are about 10% smaller in the PMPD sample. For the Black students (panel B), however, the standard errors are essentially the same or slightly larger in the PMPD sample. This finding is unexpected because Hill, Rubin, and Thomas (2000) predicted that analyses of subgroups would have more power in the PMPD design, because matching should lead to a more nearly equal representation of subgroups in the treatment and control groups. In addition, the covariate balance is less good for the Black students in the PMPD than in the randomized blocks if we compute Z scores for the differences between treatments and controls for the baseline covariates. One possibility is that because Black students could have been matched to non-Black students in the PMPD design, this subsample was actually less balanced than the subsample of Blacks from the more conventional randomized block design.

The increase in power from the PMPD design is relatively modest, even in the full sample. Of much more practical importance is the fact that the complexity of the PMPD design caused Mathematica to initially calculate incorrect baseline sample weights. The baseline weights were designed to make the follow-up sample representative of all eligible applicants for vouchers. The initial weights assigned much too little importance to the control sample in the PMPD block. Because these families represented many unselected controls, they should have been weighted heavily. The revised weights increased the weight on the PMPD controls by 620%. In contrast, the weight increased by just 18% for the rest of the sample.

This error had grave consequences for the initial inferences and analyses of the data. First, using the faulty weights, Peterson, Myers, and Howell (1998) reported highly statistically significant differences in *baseline* test scores between the treatment and control groups, with control group members scoring higher on both the math and reading exams. After the mistake was discovered and the weights were revised, however, the baseline differences became small and statistically insignificant. (We note, however, that if we limit attention to the students from single-children families that Barnard et al. use, and use the more complete baseline test score data, then there is

a statistically significant difference in baseline reading scores between treatments and controls, even with the revised baseline weights. If we pool together students from single-child and multichild families, then the baseline difference is insignificant. This is another reason why we think it is more appropriate to use the broader sample that includes children from all families.)

Second, because of the inaccurate inference that there was a difference in baseline ability between treatments and controls, Mathematica's researchers were discouraged from analyzing (or at least from presenting) results that did not condition on baseline scores. This is unfortunate, because conditioning on baseline scores caused the researchers to drop from the sample all the students who were initially in kindergarten (because these students were not given baseline tests) and 11% of students initially in grades 1–4. Including students with missing baseline scores increases the sample by more than 40%, and expands the population to which the results can be generalized. As explained later and in an article by Krueger and Zhu (2003), qualitatively different results are found if students with missing baseline scores are included in the sample. Because assignment to receive a voucher was random within lottery strata, estimates that do not condition on baseline scores are nonetheless unbiased. Moreover, it is inefficient to exclude students with missing baseline scores (most of whom had follow-up test scores), and such a sample restriction can potentially cause sample selection bias.

Of lesser importance is the fact that the PMPD design also complicates the calculation of standard errors. The matching of treatments and controls on selected covariates creates a dependence between paired observations. Moreover, if the propensity score model is misspecified, then the equation error also creates a dependence between observations. Researchers have not taken this dependence into account in the calculation of standard errors. We suspect, however, that this is likely to cause only a small understatement of the standard errors, because the covariates do not account for much of the residual variance of test scores, and because the PMPD sample is only about half of the overall sample. (We tried to compute the propensity score to adjust the standard errors ourselves, but were unable to replicate the PMPD model because it was not described in sufficient detail and because the computer programs are proprietary. This also prevented us from replicating estimates in Table 4 of Barnard et al.)

The problems created by the complicated experimental design, particularly concerning the weights, lead us to reiterate the advice of Cochran and Cox (1957): "A good working rule is to use the simplest experimental design that meets the needs of the occasion." It seems to us that in this case, the PMPD design introduced unnecessary complexity that inadvertently led to a consequential mistake in the computation of weights, with very little improvement in efficiency. This was not the fault of the architects of the PMPD design, who did not compute the weights, but it was related to the complexity of the design. Simplicity and transparency—in designs, methods, and procedures—can help avoid mistakes down the road, which are almost inevitable in large-scale empirical studies. Indeed, Mathematica recently informed us that they still do not have the baseline weights exactly correct, 5 years after random assignment.

2. INTENT-TO-TREAT ESTIMATION RESULTS

Barnard et al. devote much attention to addressing potential problems caused by missing data for the sample of students in grades 1–4 at baseline; this follows the practice of earlier reports by Mathematica, which restricted the analysis of student outcomes (but not parental responses such as satisfaction) to those enrolled in grades 1–4 at baseline. In our opinion, a much more important substantive problem results from the exclusion of students in the kindergarten cohort, who were categorically dropped because they were not given baseline tests. Because assignment to treatment status was random (within strata), a simple comparison of means between treatments and controls without conditioning on baseline scores provides an unbiased estimate of the average treatment effect. Moreover, as Barnard et al. and others show, treatments and controls were well balanced in terms of baseline characteristics, so there is no reason to suspect that random assignment was somehow corrupted.

Table 2 presents regression estimates of the ITT effect using various samples. Because we cannot replicate the Bayesian estimates without knowing the propensity score, we present conventional ITT estimates. For each entry in the first column, we regressed the test score percentile rank on a voucher offer dummy, 30 dummies indicating lottery strata (block × family size × high/low school), and baseline test scores. The sample is limited to those with baseline test data. Our results differ from the ITT estimates in Table 5 of Barnard et al., because we use the revised weights and a more comprehensive sample that also includes students from multichild families. Barnard et al. report that their Bayesian estimates are "generally more stable" than the conventional ITT estimates, "which in some cases are not even credible," but the extreme ITT estimates that they refer to for 4th graders from high-achieving schools are based on only 15 students. For the samples that pool students across grades, the results of the conventional ITT estimates are a priori credible and probably not statistically different from their Bayesian estimates.

In the second column we use the same sample as in column 1, but omit the baseline test score from the model. In the third column we expand the sample to include those with missing baseline scores. In the fourth column we continue to include students with missing baseline scores, but restrict the sample to those initially in low-achieving public schools.

We focus mainly on the results for Black students and students from low-achieving public schools, because the effect of offering a voucher on either reading or math scores for the overall sample is always statistically insignificant, and because most public policy attention has focused on these two groups.

Although it has not been made explicit in previous studies of these data, Black students have been defined as children with a non-Hispanic, Black/African-American mother. We use this definition and a broader one to probe the sensitivity of the results.

Omitting baseline scores has little qualitative effect on the estimates when the same sample is used (compare columns 1 and 2); the coefficient typically changes by no more than half a standard error. For year 3 scores for Black students, for example, the treatment effect on the composite score is 5.2 points ($t = 3.2$) when controlling for baseline scores and 5.0 points ($t = 2.6$) when not controlling. This stability is expected in a

Table 2. Estimated Treatment Effects, With and Without Controlling for Baseline Scores Coefficient on Voucher Dummy From Test Score Regression for Various Samples

Test	Sample	Subsample with baseline scores controls for baseline scores	Subsample with baseline scores omits baseline scores	Full sample omits baseline scores	Subsample applicant school: Low omits baseline scores
First follow-up test					
Composite	Overall	1.27	.42	−.33	−.60
		(.96)	(1.28)	(1.06)	(1.06)
	Mother Black, Non-Hispanic	**4.31**	**3.64**	2.66	1.75
		(1.28)	**(1.73)**	(1.42)	(1.57)
	Either parent Black, Non-Hispanic	**3.55**	2.52	1.38	.53
		(1.24)	(1.73)	(1.42)	(1.55)
Reading	Overall	1.11	.03	−1.13	−1.17
		(1.04)	(1.36)	(1.17)	(1.18)
	Mother Black, Non-Hispanic	**3.42**	2.60	1.38	.93
		(1.59)	(1.98)	(1.74)	(1.84)
	Either parent Black, Non-Hispanic	2.68	1.49	−.09	−.53
		(1.50)	(1.97)	(1.71)	(1.81)
Math	Overall	1.44	.80	.47	−.03
		(1.17)	(1.45)	(1.17)	(1.15)
	Mother Black, Non-Hispanic	**5.28**	**4.81**	**4.01**	2.68
		(1.52)	**(1.93)**	**(1.54)**	(1.63)
	Either parent Black, Non-Hispanic	**4.42**	3.54	2.86	1.59
		(1.48)	(1.90)	(1.51)	(1.63)
Third follow-up test					
Composite	Overall	.90	.36	−.38	.10
		(1.14)	(1.40)	(1.19)	(1.16)
	Mother Black, Non-Hispanic	**5.24**	**5.03**	2.65	3.09
		(1.63)	**(1.92)**	(1.68)	(1.67)
	Either parent Black, Non-Hispanic	**4.70**	**4.27**	1.87	1.90
		(1.60)	**(1.92)**	(1.68)	(1.67)
Reading	Overall	.25	−.42	−1.00	−.08
		(1.26)	(1.51)	(1.25)	(1.24)
	Mother Black, Non-Hispanic	3.64	3.22	1.25	2.06
		(1.87)	(2.19)	(1.83)	(1.83)
	Either parent Black, Non-Hispanic	3.37	2.75	.76	1.23
		(1.86)	(2.18)	(1.83)	(1.82)
Math	Overall	1.54	1.15	.24	.29
		(1.30)	(1.53)	(1.33)	(1.28)
	Mother Black, Non-Hispanic	**6.84**	**6.84**	**4.05**	**4.13**
		(1.86)	**(2.09)**	**(1.86)**	**(1.85)**
	Either parent Black, Non-Hispanic	**6.04**	**5.78**	2.97	2.58
		(1.83)	**(2.09)**	(1.85)	(1.82)

NOTE: Standard errors are in parentheses. Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for 30 randomization strata dummies; models in the first column also control for baseline math and reading test scores.
Bootstrap standard errors are robust to correlation in residuals among students in the same family. Bold font indicates that the absolute t ratio exceeds 1.96.
Sample sizes in year 1 for subsample with baseline scores/or without are overall, 1,455/2,080; mother black, 623/883; either parent black, 684/968.
Sample sizes in year 3 for subsample with baseline scores/or without are overall, 1,250/1,801; mother black, 519/733; either parent black, 572/807.
Sample sizes in year 1 for subsample from low schools without baseline scores are overall, 1,802; mother black, 770; either parent black, 843.
Sample sizes in year 3 for subsample from low schools without baseline scores are overall, 1,569; mother black, 647; either parent black, 712.

randomized experiment. When the sample is expanded to include those with missing baseline scores in column 3, however, the ITT effect falls almost in half in year 3 for Black students, to 2.65 points with a t ratio of 1.58. The effect on the math score is statistically significant, as Barnard et al. emphasize. Qualitatively similar results are found if baseline covariates, such as mother's education and family income, are included as regressors (see Krueger and Zhu 2003).

The classification of Black students used in the study is somewhat idiosyncratic. Race and ethnicity were collected in a single question in the parental survey—contrary to the OMB guidelines for government surveys—so it is impossible to identify Blacks of Hispanic origin (of which there are a large number in New York). In addition, the survey did not directly ask about the child's race or ethnicity, so children's race must be inferred from parents' race/ethnicity. A broader definition of Black students—and one that probably comes closer to the definition used in government surveys and most other educational research—would also include those who have a Black father in the sample of Black students. According to the 1990 Census, race was reported as Black for 85% of the children with a Black father and a Hispanic mother (the overwhelming non-Black portion of the sample) in the New York metropolitan region.

Hence we also present results for a sample in which either parent is listed as "Black, non-Hispanic." This increases the sample of Black students by about 10%, and the results are even weaker for this sample. For example, the ITT estimate using the more comprehensive sample of Black students enrolled in grades K–4 at baseline is 1.87 points ($t = 1.11$) on the third-year follow-up composite test, although even with this sample

the effect on math is significant at the .10 level in year 1, as Barnard et al. have found, and almost significant in year 3. So we conclude that the effect of the opportunity to use a private school voucher on the composite score for the most comprehensive sample of Black students is insignificantly different from 0, although it is possible that there was initially a small beneficial effect on the math score for Black students.

In the final column we present results for the subsample of students originally enrolled in schools with average test scores below the median score in New York City. In each case, the results for this subsample are quite similar to those for the full sample, and a formal test of the null hypothesis that students from low- and high-achieving schools benefit equally from vouchers is never close to rejecting. It also appears very unlikely that the differences between the treatment effects for applicants from low- and high-achieving schools in Barnard et al.'s Table 4 are statistically significant either.

3. WHAT WAS BROKEN?

We agree with Barnard et al. that the experiment was broken in the sense that attrition and missing data were common. Previous analyses were also strained, if not broken, for their neglect of the cohort of students originally in kindergarten who were in third grade at the end of the experiment and whose follow-up test scores were ignored. Including these students qualitatively alters the results for Black students. The experiment was also broken in the sense that years passed before correct baseline weights were computed.

We disagree, however, with the interpretation that the experiment was broken because compliance was less than 100%. This depends on the question that one is interested in answering. Because most interest in the experiment for policy purposes centers on the impact of *offering* vouchers on achievement—not on compelling students to use vouchers—we think the ITT estimates, which reflect inevitable partial usage of vouchers, are most relevant (see also Rouse 1998; Angrist et al. 2003). Moreover, New York had a higher voucher take-up rate than the experiments in Dayton and the District of Columbia, so one could argue that the New York experiment provides an upper bound estimate of the effect of offering vouchers. On the other hand, if the goal is to use the experiment to estimate the effect of *attending* private school on achievement, then such methods as instrumental variables or those used by Barnard et al. are necessary to estimate the parameter of interest. We consider this of secondary interest in this case, however.

ADDITIONAL REFERENCES

Angrist, J., Bettinger, E., Bloom, E., King, E., and Kremer, M. (2003), "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92, 1535–1558.

Cochran, W., and Cox, G. (1957), *Experimental Designs* (2nd ed.), New York: Wiley.

Krueger, A., and Zhu, P. (2003), "Another Look at the New York City School Voucher Experiment," *American Behavioral Scientist*, forthcoming. Also available from Industrial Relation Section, Princeton University, at *http://www.irs.princeton.edu*.

Peterson, P., Myers, D., and Howell, W. (1998), "An Evaluation of the New York City Scholarships Program: The First Year," Mathematica Policy Research, Inc., available at *http://www.ksg.harvard.edu/pepg/pdf/ny1rpt.pdf*.

Rouse, C. (1998), "Private School Vouchers And Student Achievement: An Evaluation Of The Milwaukee Parental Choice Program," *The Quarterly Journal of Economics*, 13, 553–602.

Comment

Richard A. BERK and Hongquan XU

1. INTRODUCTION

The article by Barnard, Frangakis, Hill, and Rubin (hereafter BFHR) is a virtuoso performance. By applying the Neyman–Rubin model of causal effects and building on a series of important works on the estimation and interpretation of casual effects (e.g., Angrist, Imbens, and Rubin 1996), BFHR manage to extract a reasonable set of findings for a "broken" randomized experiment. The article more generally underscores the important difference between estimating relationships in a consistent manner and interpreting those estimates in causal terms; obtaining consistent estimates is only part of the enterprise. Finally, the article emphasizes that assumptions made so that proper estimates may be obtained are not just technical moves of convenience. Rather, they are statements about how the empirical world is supposed to work. As such, they need to be examined concretely with reference to the phenomena being studied, evaluated empirically whenever possible, and, at a minimum, subjected to sensitivity tests.

As battle-tested academics, we can certainly quibble here and there about some aspects of the article. For example, might it not have made sense to construct priors from the educators and economists who were responsible for designing the intervention? Moreover, might it not have made sense to interpret the results using a yardstick of treatment effects that are large enough to matter in practical terms? Nonetheless, we doubt that overall we could have done as well as BFHR did. Moreover, most of our concerns depend on features of the data that cannot be known from a distance. We would have had to carefully examine the data ourselves. As a result, we focus on why a virtuoso performance was needed to begin with. Why was this trip necessary?

Richard A. Berk is Professor and Hongquan Xu is Assistant Professor, Department of Statistics, University of California, Los Angeles, CA 90095 (E-mail: *berk@stat.ucla.edu*).

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Applications and Case Studies
DOI 10.1198/016214503000107

2. NONCOMPLIANCE

The implementation of randomized field experiments, as desirable as they are, invites any number of well-known problems (Berk 1990). One of these problems is that human subjects often do not do what they are told, even when they say they will. There is recent scholarly literature on noncompliance (Metry and Meyer 1999; Johnson 2000) and a citation trail leading back to evaluations of the social programs of the War on Poverty (Havemen 1977). Some noncompliance can be expected except when there is unusual control over the study subjects. Our recent randomized trial of the inmate classification system used by the California Department of Corrections is one example (Berk, Ladd, Graziano, and Baek 2003).

An obvious question follows: Why does it seem that the individuals who designed this experiment were caught by surprise? As best we can tell, little was done to reduce noncompliance, and, more important for the analyses of BFHR, apparently almost no data were collected on factors that might be explicitly related to noncompliance (Hill, Rubin, and Thomas 2000).

Without knowing far more about the design stages of the experiment, we find it difficult to be very specific. But from the literatures on compliance and school choice, the following are probably illustrative indicators relevant for parents who would not be inclined to use a school voucher; for other forms of noncompliance, other measures would be needed:

1. Whether the adults in the household are employed outside the home during the day
2. Whether there is any child care at the private schools before classes begin and after they end for children of employed parents
3. A household's ability to pay for private school beyond the value of the voucher
4. Travel time to the nearest private schools
5. Availability of transportation to those schools
6. Religious preference of the household and religious affiliation of the private school.

From these examples, and others that individuals closer to the study would know about, one can imagine a variety of measures to reduce noncompliance. For instance, key problems for any working parent are child care before and after school, and round-trip school transportation. These might be addressed directly with supplemental support. If no such resources could be obtained, then the study design might be altered to screen out before random assignment households for whom compliance was likely to be problematic.

3. MISSING DATA

No doubt, BFHR know more about the subtleties of the missing data than we do. Nevertheless, the same issues arise for missing data that arise for noncompliance. Surely, problems with missing data could have been anticipated. Factors related to nonresponse to particular items could have been explicitly measured and prevention strategies perhaps could have been deployed.

4. DESIGN EFFICIENCY

A carefully designed experiment not only provides a solid basis for statistical inference, but also gives credible evidence for causation. The New York School Choice Scholarships Program (NYSCSP) was apparently the first good example of a randomized experiment to examine the potential benefits of vouchers for private schools. Offering a scholarship (treatment) randomly through a lottery can (1) provide a socially acceptable way to ration scarce resources, (2) protect against undesirable selection bias and (3) produce balance between the observed and unobserved variables, which may affect the response. In addition to the randomization and blocking, NYSCSP implemented a PMPD (propensity matched pairs design) to choose a control group from a large candidate pool and to balance many covariates explicitly. Table 2 of BFHR shows that various background variables were balanced properly between the treatment group and the control group.

But, might it have been possible to do a bit better? BFHR did not mention that the sample size of the applicant's school was not balanced, although winners from high and low schools were balanced between the treatments and the controls. An ideal design would have had equal winners from "bad" (or low) and "good" (or high) public schools. In the NYSCSP, 85% of the winners were from bad schools and 15% were from good schools, to achieve the social goals of the School Choice Scholarships Foundation (Hill et al. 2000, p. 158). Bad schools are defined as "those for which the average test scores were below the median test scores for the city" (Hill et al. 2000, p. 158). A direct consequence is that the estimates from the good schools have much larger variation than that from the bad schools, which was evident in BFHR's Tables 4–6. The confidence intervals for good schools are wider.

Instead of classifying schools into two classes (bad and good), it might have been helpful to classify them into four classes: very bad, bad, good, and very good. To estimate the effect of applicant's school (assuming linear effects), an optimal design would select half of the winners to the very bad school, other half to the very good school, and none to the two middle schools. Avoiding the middle schools enlarges the distance between bad and good schools and also reduces the source variation in the bad and good schools, and thus reduces the variation in the estimates. To be more consistent with the goals of the School Choice Scholarships Foundation, an alternative design might take 75% winners from very bad schools, 10% from bad schools, 5% from good schools, and 10% from very good schools. This design would have a higher design efficiency than the original design especially because the applicant's school was the most important design variable after family size (Hill et al. 2000, p. 160). Of course, people on the ground at the time may have considered these options and rejected them. Our point is that with better planning, it might have been possible to squeeze more efficiency out of the study design. Such an outcome would have made the job undertaken by BFHR a bit easier.

5. CONCLUSIONS

Scientific problems often provide a fertile ground for important statistical developments. BFHR's article and the literature on which they depend are excellent illustrations. However, from

the perspective of public policy, the goal of program evaluation is to obtain usefully precise estimates of program impacts while using the most robust means of analysis available. Simplicity is also important, both for data analysts who must do the work and policy makers who must interpret the results. It follows that large investments at the front end of an evaluation—in research design, measurement, and study integrity—are essential. Virtuoso statistical analyses can certainly help when these investments are insufficient or when they fail, but they must not be seen as an inexpensive alternative to doing the study well to begin with.

ADDITIONAL REFERENCES

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 441–472.

Berk, R. A. (1990), "What Your Mother Never Told You About Randomized Field Experiments," in *Community-Based Care of People With AIDS: Developing a Research Agenda*, AHCPR Conference Proceedings, Washington, DC: U.S. Department of Health and Human Services.

Berk, R. A., Ladd, H., Graziano, H., and Baek, J. (2003), "A Randomized Experiment Testing Inmate Classification Systems," *Journal of Criminology and Public Policy*, 2, 215–242.

Havemen, R. H. (ed.) (1977), *A Decade of Federal Antipoverty Programs: Achievements, Failures, and Lessons*, New York: Academic Press.

Hill, J. L., Rubin, D. B., and Thomas, N. (2000), "The Design of the New York School Choice Scholarships Program Evaluation," in *Research Design: Donald Campbell's Legacy* (Vol. II), ed. L. Bickman, Thousand Oaks, CA: Sage.

Johnson, S. B. (2000), "Compliance Behavior in Clinical Trials: Error or Opportunity?," in *Promoting Adherence to Medical Treatment in Chronic Childhood Illness: Concepts, Methods, and Interventions*, ed. D. Drotar Mahwah, NJ, Lawrence Erlbaum.

Metry, J., and Meyer, U. A. (eds.) (1999), *Drug Regimen Compliance: Issues in Clinical Trials and Patient Management*, New York: Wiley.

Rejoinder

John BARNARD, Constantine E. FRANGAKIS, Jennifer L. HILL, and Donald B. RUBIN

We thank the editorial board and the thoughtful group of discussants that they arranged to comment on our article. The discussants have raised many salient issues and offered useful extensions of our work. We use this rejoinder primarily to address a few points of contention.

Muthen, Jo, and Brown

Muthen, Jo, and Brown (henceforth MJB) provide an enlightening discussion of an alternate and potentially complementary approach to treatment effect estimation. The growth models that they discuss are for data more extensive than what we analyzed and so are beyond the scope of our article. Nevertheless, MJB propose an interesting idea for looking for interactions in longitudinal experimental data, and we look forward to future applications of the method to appropriate datasets.

Although models that incorporate both our compliance principal strata as well as latent growth trajectories are theoretically possible (assuming multiple years of outcome data) and could potentially yield an extremely rich description of the types of treatment effects manifested, the more time points that exist, the more complicated the compliance patterns can become (at least without further assumptions), and it may be difficult to find a dataset rich enough to support all of the interactions. Given a choice, we feel that it is probably more beneficial to handle real observed issues than to search for interactions with latent trajectories, but certainly it would always be nice to be able to do both. Of course, such a search for interactions in trajectories is more beneficial when there is enough prior knowledge about the scientific background to include additional structural assumptions that can support robust estimation.

MJB make an interesting point about the exploration of potential violations of exclusion for never-takers. Such analyses have been done in past work by some of the authors (Hirano et al. 2000; Frangakis, Rubin, and Zhou 2002), but was not pursued for our study, because the robustness of such methods is still an issue in examples as complex as this one.

Regarding the trade-offs between the joint latent ignorability and compound exclusion assumptions versus the joint assumptions of standard ignorability and the standard exclusion restriction, this topic is up for debate. We would like to point out, however, that these standard structural assumptions come with their own level of arbitrariness. These issues have been explored by Jo (2002) and Mealli and Rubin (2002, 2003).

We thank MJB for providing a thought-provoking discussion of our model and assumptions, as well as ample fodder for additional directions that can be explored in this arena. Clearly the general category of work has stimulated much new and exciting modeling.

Krueger and Zhu

Krueger and Zhu (henceforth KZ) focus primarily on the underlying study and analyses that either we did not do or used data we did not have. Clearly, they are interested in the substantive issues, which is admirable, but we cannot be responsible for analyses that we neither did nor were consulted on, and are not prepared to comment on what analyses we might have undertaken with data that we did not have.

KZ state that "there is no substitute for probing the definition and concepts that underlie that data." Although we are not entirely sure what this refers to in particular, if it means that good science is more important than fancy techniques, then we certainly agree.

Another comment by KZ is that "it is desirable to use the most recent, comprehensive data, for the widest sample possible." We agree that it is desirable to have access to the most comprehensive data available, but it is not necessary to use all of these data in any given analysis. Here we use only first-year data, because these were the only data available to us at the

Journal of the American Statistical Association
June 2003, Vol. 98, No. 462, Applications and Case Studies
DOI 10.1198/016214503000116

time that we were building the model. We used only single-child families, as explained in the article, because, regrettably, no compliance data were collected for individual children in the multichild families.

The exclusion of kindergarten children could be considered more of a judgment call regarding the quality of the pretest scores, not only for efficiency, but also for the reasonableness of our assumptions about noncompliance and missing data. Nothing in theory would prevent us from including these children and omitting these extra covariates.

These types of trade-offs (more data vs. more appropriate data) are common in statistical analyses and the benefits are well known, for example, for avoiding model extrapolation or when some subsets of data do not mean the same thing as others. In fact, Krueger "ignored data" in his article on minimum wage with David Card (Card and Krueger 1994), by limiting analyses to restaurants from a similar geographic region in the hope that employees and employers across the two comparison states would be faced with similar economic conditions, rather than looking at all restaurants in these states. This is a form of implicit matching, which we believe represented a good choice. Surely, there are times when focusing on subsets of data produces more reliable results.

Furthermore, we would like to point out that KZ's assertion that "because assignment to treatment status was random (within strata), a simple comparison of means between treatments and controls without conditioning on baseline scores provides an unbiased estimate of the average treatment effect" is simply false, because there are missing outcomes. Even if there is perfect compliance, a simple comparison of means across treatment groups when there are missing outcomes will not lead in general to an unbiased estimate of the treatment effect. Assumptions need to be made about the missing-data process.

When noncompliance exists in addition to the missing outcome data, Frangakis and Rubin (1999) have shown that estimation even of the ITT effect requires additional assumptions. Moreover, the ITT estimate, generally considered a conservative estimate of the treatment effect, can actually be anticonservative in this scenario. Finally, contrary to what KZ seem to imply in their final paragraph, instrumental variables methods alone are not sufficient to address the problems of both noncompliance and missing data.

KZ advocate "simplicity and transparency." We wholeheartedly agree, but disagree with the examples they use to defend their stance. Their first argument is that the PMPD did not yield strong efficiency gains and that we did not account for the PMPD in the calculation of standard errors. It may well be that precision gains were relatively modest; however, the analyses that they present are by no means conclusive in this regard. First, as with KZ's other ITT analyses, we have no clue as to how missing outcome data are handled. In addition, the PMPD was also used to balance compliance and missingness. It is unclear, however, how the gains in precision due to matching the missingness and noncompliance are displayed, if at all, in KZ's results. Finally, KZ did not account for the experimental design when estimating their results. Failure to account properly for the design typically results in estimates that overstate the variances of treatment effect estimates. In this design, it is the correlation between matched pair units that leads the sampling variances to be smaller than their estimates.

Even if precision gains were truly modest, they of course were unknown to us at the time we debated the use of the PMPD design. KZ's argument is like the complaint of a cross-country motorist who experiences no car problems during his long drive and then bemoans the money wasted on a spare tire purchased before the trip began. We prefer to avoid potential problems by using careful design.

Importantly, KZ's criticism that we ignored this matching in the analysis is incorrect. The matching was done based on the estimated propensity score, so it follows from Rubin (1978) that assignment is ignorable conditionally given the covariates that were used to estimate it. Thus the Bayesian analysis that we conduct that conditions on the estimated propensity score is an approximately valid Bayesian analysis with respect to the matched assignment, assuming that our modeled relationship between the outcomes and the estimated propensity score captures the major features of the relationship to the covariates.

KZ's second argument is that use of the PMPD "led to grave consequences" because sampling weights were miscalculated by MPR. It seems to us odd to maintain, as KZ do, that the PMPD "caused" this error, nor is it clear that use of another design would have avoided the error. In any case, we have rerun our analyses, this time generalizing to the population of which the children in our analyses are representative (by setting constant weights in step 3 of Sec. 8.1.1). For this population, all results are very similar to our original results. These results are posted as Appendix C at *http://biosun01.biostat.jhsph.edu/~cfrangak/papers/sc*.

A third argument revolves around the protracted debate between use of ITT versus treatment-targeted measures such as CACE. We addressed both ITT and CACE, which allows the reader to focus on one or the other or both, depending on which effect is expected to generalize to another environment: the ITT, or "sociological" effect of accepting or not accepting an offered voucher, or the CACE, or "structural" or "scientific" effect of using a voucher. One could argue that the effect of attending (CACE) is the more generalizable, whereas the sociological effect (ITT) is more easily influenced, for instance by "hype" about how well the program works. That is, the ITT estimate will not reflect the "effect of offering" if the next time the program is offered the compliance rates change (which seems likely!). We prefer to examine both to provide a more comprehensive picture.

In sum, we agree that simplicity and transparency are important; this is the major motivation for our detailed description of our assumptions in terms of readily understandable concepts rather than commonly invoked and less intuitive conditions involving correlated error terms. However, this desire for simplicity must be balanced with the prudence of guarding against potential problems, through careful design.

Although we agree in theory with much of what KZ espouse, some of their criticisms seem misguided. Moreover, their presentation of analyses with incomplete exposition regarding missing-data complications makes these analyses difficult for us to evaluate properly.

Berk and Xu

Berk and Xu (henceforth BX) raise several interesting questions about the way in which the evaluation was designed and implemented. Many of these are actually addressed in a preceding and lengthier article by us on this topic (Barnard et al. 2002). That article describes, for instance, MPR's efforts to reduce missing data, without which the missingness rates would surely have been much higher. It also describes special challenges faced in this study. But in any study with human subjects, it is difficult to completely eliminate missing data, particularly if one desires more interesting information than can be obtained from administrative data. One of the challenges for the collection of outcome test scores was the fact that children needed to be tested outside of the public school setting, which allowed the use of a standardized test not used by the public school system, thereby greatly reducing the possibly of bias resulting from teachers "teaching to the test." Consequently, parents had to bring their children to testing areas on weekends—a formidable hurdle for some even when monetary incentives were used.

An important consideration with this study was that the evaluators took the initiative to investigate a program that was sponsored by an organization (SCSF) whose primary goal was to provide scholarships, not to do research. We feel that MPR did an admirable job convincing the SCSF to allow the study to be evaluated and raising money for this purpose, but this also meant that the evaluators could not control every aspect of the program. For example, the SCSF originally only wanted to make the program available to children from the lower-test-score schools, and the evaluators had to convince them to allow also a small percentage from those in higher-test-score schools, to increase the generalizability of the results. The designs that BX suggest, even though they have certain desirable statistical properties under appropriate assumptions, were not an option with this study.

Related to this issue of the alignment of the evaluators' and program administrators' goals is the issue of the potential for reducing noncompliance rates. Some effort was made in that help was provided to scholarship winners in finding an appropriate private school. Arguably, however, 100% compliance should not be the goal in such a study, given that we would not expect 100% compliance if the program were expanded and offered, say by the government, on a larger scale. Thus it may not be wise to use extreme efforts to increase compliance if such efforts would not be offered in the large-scale public case. With the current study, the opportunity exists to examine the noncompliance process for this program in this setting: take-up rates, predictors of noncompliance, and so on. To this end, some sort of subexperiment that randomizes a reasonable set of compliance incentives might be useful in similar studies in the future, if sample sizes are sufficient.

We are not convinced that we should have solicited actual prior distributions from educators and economists. In our experience, this is a confusing, difficult, and often unrewarding task. Moreover, school choice is such a politically charged and contentious issue that it seems wiser to try to use relatively diffuse and objective prior distributions.

We agree that results should be interpreted in terms of a yardstick of practical effects. We are unaware of the most appropriate such yardstick to use in this context. One side could claim that even tiny average effects matter, not only because of the cumulative effects on parts of society, but also because of possible nonadditivity of such effects, which could imply huge effects for some rare individuals and minute effects for all others. Others could argue that seemingly quite large effects have no long-run practical impact on quality of life. The results that we present provide a basic empirical starting point for these debates between people whose expertise is in this area.

We thank BX for highlighting our emphasis on the role of the assumptions in our analyses. Our hope is that we have presented the assumptions in such a way to enable readers to make their own judgment about whether or not they were likely to be satisfied, rather than only stating them technically, which would implicitly make those important decisions for the interested but nontechnical readers.

BX point out a dearth of certain types of covariates. Part of the problem here is that we presented only a small portion of the data collected, due to the fact that our model in its current form did not handle many covariates. However, there were indeed discussions between some members of our team and MPR staff during the development of the questionnaire about what types of variables could be included that would be predictive of missing data or compliance, as well as outcomes. Some of the variables proposed were discarded due to the surveyors' experience with them as subject to problems, such as high measurement error. Also, not every variable discussed was included due to a desire for a survey that would not be overly burdensome for the study families.

We raise a final small semantic question. We wonder about the propriety of the label "Neyman–Rubin model" to describe a framework for causal inference that views observational studies and randomized experiments on the same continuum, and moreover encourages Bayesian analyses of both. Neyman (1923) certainly seems to have been the first to use the formal notation of outcomes for randomization-based potential inference in randomized experiments, which was a truly major advance, but neither he nor anyone else to the best of our knowledge used this formal notation to define causal effects in observational studies until Rubin (1974, 1975), nor did he nor anyone else ever advocate Bayesian inference for the unobserved potential outcomes until Rubin (1975, 1978), which is the perspective that we have implemented here.

Overall, we agree with BX about the paramount importance of study design, and we tried to design this study carefully, certainly more so than most social science analyses of secondary datasets like the PSID or CPS. We thank BX for their thoughtful commentary; they raise some important points.

We thank all of the discussants for stimulating comments, which we believe have added emphasis to the importance of both the substantive topic of school vouchers and the development of appropriate methods to evaluate this and other such broken randomized experiments.

ADDITIONAL REFERENCES

Card, D., and Krueger, A. (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–793.

Jo, B. (2002), "Estimation of Intervention Effects With Noncompliance: Alternative Model Specifications" (with discussion), *Journal of Educational and Behavioral Statistics*, 27, 385–420.

Mealli, F., and Rubin, D. B. (2002), Discussion of "Estimation of Intervention Effects With Noncompliance: Alternative Model Specifications," by B. Jo, *Journal of Educational and Behavioral Statistics*, 27, 411–415.

——— (2003), Assumptions Allowing the Estimation of Direct Causal Effects: Discussion of "Healthy, Wealthy, and Wise? Tests for Direct Causal Paths Between Health and Socioeconomic Status" by Adams et al., *Journal of Econometrics*, 112, 79–87.

Rubin, D. B. (1975), "Bayesian Inference for Causality: The Importance of Randomization," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 233–239.