School Choice in NY City: A Bayesian Analysis of a Broken Randomized Experiment*

John Barnard Constantine Frangakis Jennifer Hill David Myers
Paul Peterson Donald B. Rubin

September 14, 1999

Abstract

The supposed decline of the U.S. educational system, including its potential causes and solutions, has been a popular topic 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. For example, educational researchers rarely are able to engage in controlled, randomized experiments. The efficacy of so-called "school choice" programs has been a particularly contentious issue. A current multi-million dollar evaluation of the New York School Choice Scholarship Program (NYSCSP) endeavors to shed some light on this issue. This study can be favorably contrasted with other school choice evaluations in terms of the consideration that went into the randomized experimental design (a completely new design, the Propensity Matched Pairs Design, is being implemented) and the rigorous data collection and compliance-encouraging efforts. In fact, this study benefits from the authors' previous experiences with the Milwaukee Parental Choice Program, which, although randomized, was relatively poorly implemented as an experiment.

At first glance, it would appear that the evaluation of the NYSCSP could proceed without undue statistical complexity. Unfortunately, this program evaluation, as is common in studies with human subjects, suffers from unintended complications. The first complication is non-compliance. Approximately 25% of children who were awarded scholarships decided not to use them. The second complication is missing data: some guardians failed to complete fully survey information; some children were too young to take pre-tests; some children failed to show up for post-tests. Levels of missing data range approximately from 3 to 50% across variables. Work by Frangakis and Rubin (1999) has revealed the severe threats to valid estimates of experimental effects that can exist in the presence of non-compliance and missing data, even for estimation of simple intention-to-treat effects.

The technology we use to proceed with analyses of longitudinal data from a randomized experiment suffering from missing data and non-compliance involves the creation of multiple imputations, for both missing outcomes and missing true compliance statuses, as well as for missing covariates for some analyses, using Bayesian models. The fitting of Bayesian models to such data requires MCMC methods for missing data. Our Bayesian approach allows for analyses that rely on fewer assumptions than standard approaches.

*This draft has not been fully read by all co-authors, and so primary responsibility for errors and omissions lies with John Barnard, Jennifer Hill and Donald Rubin.

1 Prologue

Every day policy decisions are made that may have a great impact on our lives based on quantitative "analyses" of social science data. Rigorous mathematical statisticians are sometimes wary of participating in social science analyses because social science data sets are nearly always messy relative to those in the physical or biological sciences even when statisticians are involved in the design of the study. Human subjects are capricious, randomized experiments can rarely be performed, and the measures are often only loosely tied to the phenomena of interest, as well as being intrinsically noisy. However, we believe that it is incumbent on the statistician who choses to plow these rocky fields to do so with the same care and energy as statisticians who chose to plow other fields.

The Bayesian paradigm, because of its flexibility, is a powerful way to conceptualize how to approach such messy problems from the design through the analysis stage. Using this paradigm as a guide does not necessarily imply performing formal Bayes calculations at each step because these might be impossibly demanding in the time frame or with available resources. However it does mean that we design a study with the eventual Bayesian analyses in mind, where design here is broadly defined to include not only the plan for assigning treatments to individuals but also evaluation issues such as the plan for what types of data will collected and in what manner. We want to design to minimize problems at the end without being blind to the eventual complications that will nearly certainly arise Rather, we optimally would like to structure these complications within our Bayesianly inspired template for the study and its data.

Of particular importance, knowing which issues create most problems for our ultimate Bayesian analysis and which variables would be most beneficial to it, helps guide our design. In fact, much of the benefits of practical importance in a study such as this arise through the design: deciding how to minimize the complications and if these complications can be incorporated into the analyses. If there are no complications the payoff to being Bayesian is typically small. In our setting, the evaluation of a program which may have major impact on lives of our children, there will be an emphasis on these design aspects but in the context of this conceptualization that is oriented toward

being Bayesian.

In this way, this application may stand in contrast to many Bayesian applications which often focus on analyses of existing datasets thereby showcasing clever modeling and computation but neglecting issues of how the data were obtained, or how the data collection was influence by the Bayesian analyses to be conducted and used to draw practical conclusions. We present as equally important aspects of the study: (1) our assessment of the most important complications involved(non-compliance with treatment assignment, missing outcomes, and missing covariates), and (2) our attempts to minimize these complications and to accommodate eventual incorporation of these complications into the analysis (for instance inclusion of survey questions intended to help in the modeling of these complications).

This application may also stand in contrast to some Bayesian applications because we compare our methodology to standard software and techniques commonly used in similar problems.

We don't feel that this draft represents a completely satisfactory job of simultaneously handling all the complications. As an example, we don't model the multivariate nature of the outcomes.

2 Introduction

Over the past few years, the school choice movement has escalated. Congress and many state legislatures have considered school voucher proposals that enable families, particularly low-income families, to choose among a wide range of schools, public and private, religious and secular. In 1990 the Wisconsin legislature enacted a pilot program that gave public students access to secular private schools in the City of Milwaukee; then in 1996 the legislature expanded this program to include religious schools. After surviving a constitutional challenge, the program went into effect in the fall of 1998. A similar program in Cleveland, enacted by the Ohio legislature, began its third year of operation in the fall of 1998. At the federal level, a pilot program for the District of Columbia received congressional approval in the summer of 1998, but was vetoed by

President Clinton.

Special interest groups, political leaders and policy analysts on all sides of the ideological spectrum have offered strong arguments concerning the continuation and/or expansion of these school choice programs. The debate is charged, and we remain far from any kind of consensus. Supporters of school choice assert that low-income, inner-city children learn more in private schools; critics retort that any perceived learning gains in private schools are due to the selected nature of private-school families. Proponents insist that families develop closer communications with schools they themselves choose; critics reply that when choices are available, mismatches often occur and private schools expel problem students, adding to the educational instability of children from low-income, inner-city families. Champions of choice suggest that a more orderly educational climate in private schools enhances learning opportunities, whereas opponents declare that private schools select out the "best and the brightest," leaving behind the most disadvantaged. Voucher advocates argue that choice fosters racial and ethnic integration; critics, meanwhile, insist that private schools balkanize the population into racially and ethnically homogeneous educational environments[1]

Few of these disputes have been resolved, in part because high-quality information about the effects of school-choice programs is in short supply. Although many published studies compare public and private schools, they have been consistently criticized for comparing dissimilar populations. Even when statistical adjustments are made for background characteristics, it remains unclear whether findings reflect actual differences between public and private schools or simply differences in the kinds of students and families attending them[2].

Though this problem has plagued educational research for years, it is not insurmount-

[1]Recent works making a case for school choice include Brandl (1998); Coulson (forthcoming); Cobb (1992); and Bonsteel and Bonilla (1997). A collection of essays that report mainly positive school-choice effects are to be found in Peterson and Hassel (1998). Works which critique school choice include Ascher, Fruchter, and Berne (1996); Carnegie Foundation for the Advancement of Teaching (1992); Gutmann (1987); Levin (1998); Fuller and Elmore (1996); Rasell and Rothstein (1993); Cookson (1994).

[2]Major studies finding positive educational benefits from attending private schools include Coleman, Hoffer, and Kilgore (1982); Chubb and Moe (1990); Derek (1997). Critiques of these studies have been prepared by Goldberger and Cain (1982); Wilms (1985).

able. The best solution is to implement controlled randomized designs. Randomized experiments, though standard in other fields, have only recently found their way into educational studies, such as the Tennessee Star experiment, which found that smaller classes have positive effects among students in kindergarten and first grade (Mosteller 1995). Until now, however, randomized designs have not been carefully used to study the validity of competing claims about school choice.

In this article, we describe a case study of a randomized experiment conducted in New York City made possible by the School Choice Scholarships Foundation (SCSF), a privately-funded school choice program. The program provides the first opportunity to estimate the impacts of a school choice pilot program that has the following characteristics: a lottery that allocates scholarships randomly to applicants, which has been administered by an independent evaluation team that can guarantee its integrity; baseline data on student test performance and family background characteristics collect from students and their families prior to the lottery; data on a broad range of characteristics collected from as much as 83 percent of the test group and control group one year later. Because it has these qualities, the SCSF is an ideal laboratory for studying the effects of school choice on outcomes such as parental satisfaction, parental involvement, school mobility, racial integration and, perhaps most noteworthy, student achievement.

The school choice initiative in New York is described in Section 3 followed by study objectives and implementation in Sections 4 and 5. The innovative randomized design developed for this study is presented in detail in Section 6. Section 7 introduces the template of the broken randomized experiment and the corresponding notation is given in Section 11. The model is described in Sections 12; technical details of the computations are reserved for the Appendix. Results of the analysis are discussed in Section 13.

3 School Choice Scholarships Foundation Program (SCSFP)

In February 1997 SCSF announced that it would provide 1,300 scholarships to low-income families currently attending public schools. These scholarships were worth up to $1,400 annually, and could be used for up to three years to help pay the costs of

attending a private school, either religious or secular. SCSF received initial applications from over 20,000 students between February and late April 1997.

In order to become eligible for a scholarship, children had to be entering grades one through five, live in New York City, attend a public school at the time of application, and come from families with incomes low enough to qualify for the federal government's free school lunch program. To qualify, students and an adult member of each family had to attend verification sessions where SCSF administrators documented family income and children's public-school attendance.

Because of the large number of initial applications, it was not feasible to invite everyone to these verification sessions. To give all families an equal chance of participating, therefore, a preliminary lottery was used to determine who would be invited to a verification session. Only these families were then included in the final lottery that determined the allocation of scholarships among applicants.

The final lottery, held in mid-May 1997, was administered by Mathematica Policy Research (MPR); SCSF announced the winners. Within the guidelines established by SCSF, all applicants had an equal chance of winning the lottery. SCSF decided in advance to allocate 85 percent of the scholarships to applicants from public schools whose average test scores were less than the citywide median. Consequently, applicants from these schools, who represented about 70 percent of all applicants, were assigned a higher probability of winning a scholarship.

Subsequent to the lottery, SCSF helped families find placements in private schools. By mid-September 1997, SCSF reported that 1,168 scholarship recipients, or 75 percent of all those offered a scholarship, had successfully gained admission to some 225 private schools.

4 Objectives of the Study

The evaluation of the School Choice Scholarship Foundation (SCSF) was headed by co-principal investigators Paul Peterson and MPR (henceforth the evaluation team will be referred to solely as MPR for simplicity). The evaluation provides answers to three

questions. First, what is the impact of being offered a scholarship on student and parent outcomes? Second, what is the impact of using a scholarship (participating in the scholarship program)? That is, what is the value-added of using a scholarship over and above what families and children would do in the absence of the scholarship program (which could include either public or private school attendance)? Third, what is the impact of attending a private school on student and parent outcomes? That is, would students who attend public schools do better academically if they attended private schools? We may answer each of these questions directly by using information collected for the SCSF evaluation. Until this evaluation, no one study has addressed these three questions. Furthermore, this study can produce highly credible evidence concerning these questions because we randomly assigned families to a treatment group (offer of a scholarship) and a control group.

5 Implementation

In order to evaluate the voucher program, SCSF collected data on family demographics, parents' opinions on matters relating to their child's education, and student test scores, both before and after scholarships had been awarded; one of the conditions for participating in the program was agreement to provide confidential baseline and follow-up information. MPR also made extensive efforts to encourage cooperation with the study guidelines.

5.1 Issues in the Implementation of the SCSF Evaluation

A critical issue in the design, implementation, and analysis of a random assignment experiment, such as the evaluation of the SCSF program, concerns deviations from study protocols by families and children. We have identified four such behaviors:

1. Families assigned to the treatment group did not use a scholarship to attend a private school.
2. Families in the control group sent their children to a private school; this assumes that receipt of a scholarship and private school attendance are equivalent treat-

ments and that the range of private schools attended by the treatment group are similar to the private schools attended by the control group.

3. Families invited to attend data collection and testing sessions after the baseline survey did not show up.

4. Parents and students did not complete all items in their questionnaire, or students did not complete enough items in the standardized reading and math assessments to be given a score.

The first two of these behaviors will henceforth be referred to under the general rubric of "noncompliance," the last two as "missing data." For the SCSF evaluation, ensuring compliance with the assigned treatment was out of the control of the evaluation team. However, in other evaluations, there may be more control over noncompliance of the control group with respect to participating in program services. For example, an evaluator may work closely with program operators to ensure that they do not provide services to members of the control group. However, this cannot stop members of the control group from going out and finding similar services if they are available in the community; sometimes the services may be more or less intensive than those offered by the program being studied.

Evaluators generally have some control, however, over the amount or kinds of missing data which occur. Below, we describe the procedures used to minimize missing data.

5.2 Collection of Baseline Data

During the verification sessions at which eligibility was determined, MPR asked students to take the Iowa Test of Basic Skills (ITBS) in reading and mathematics. Students in kindergarten applying for a scholarship for first grade were exempt from the testing. Each student's performance was given a national percentile ranking. While their children were taking tests, MPR asked parents to complete questionnaires that would provide information on their satisfaction with the school their child was currently attending, their involvement in their child's education, and their background characteristics. Discussions regarding what questions to include on the baseline survey focused

not only on what types of covariates were expected to be predictive of the primary outcomes of interest, but also what might be predictive of compliance behavior and propensity towards non-response.

Although grandmothers and other relatives and guardians also accompanied children to verification sessions, in over 90 percent of the cases it was a parent who completed the questionnaire. MPR held the sessions at private schools, where students took the tests in classroom settings. In nearly all cases, private school teachers proctored the tests and were under the supervision of MPR staff. The verification sessions took place during March, April, and early May 1997 on weekends and vacation days.

5.3 Collection of Follow-Up Data in 1998

The first follow-up data collection was completed in summer 1998. MPR invited each of the 1,960 families in the treatment group and the control group to attend testing sessions. Most of the testing sessions were held on weekends during spring 1998. MPR held the testing sessions at parochial schools and parents were asked to complete a questionnaire that included many of the same items that were part of the baseline questionnaire. Students in grades 3-5 were given a questionnaire. The response rates for the first follow-up data collection are shown in Table 1. The overall response rate for the parent survey was 84 percent for the scholarship families and 80 percent for families in the control group. To achieve these high response rates, MPR used two forms of incentives. First, they offered all families in the control group a chance to win a scholarship for $1,400 for three years, but to be eligible, families and their children were required to attend a testing session and the children had to have been eligible for a scholarship when SCSF first made the scholarship awards. To preserve the integrity of the control group, we[3] randomly selected about 100 winners for the second year of scholarships. Second, each control group family that attended a testing session received an incentive of $75 on average (some were offered $50 and others were offered $100).

[3]This was actually performed by co-author Neal Thomas.

Scholarship Users	89%
Scholarship Decliners	66%
Treatment Group Total	84%
Control Group Total	80%

Table 1: Response Rates on the First Follow-Up Parent Survey

5.4 Item Nonresponse

To minimize item nonresponse in the survey questionnaires, staff at each data collection session reviewed the questionnaires for completeness as parents and students returned them at the end of the testing session. In cases where many items appeared to have been left incomplete, staff asked the parents and students to try to complete the items. If a parent or child did not understand the item, staff would work with them so that they might be able to provide a response. Sometimes, one parent would refuse to answer about the other parent if they were no longer living in the home. In Table 2, we illustrate the variability in item nonresponse rates that occurred in the baseline survey. It becomes quite clear upon reviewing these results that often there was little information concerning a child's father. For example, among the parent questionnaires, more than 35 percent of them were missing information about fathers' educational attainment and almost 60 percent were missing information about fathers' employment. In contrast, missing values were present for about seven percent of the responses concerning mothers' education and mothers' employment.

5.5 Additional Complications with the Data

Two additional complications with the data are noteworthy. The first is that no pre-test scores could be obtained for applicants in kindergarten because they were too young to take written standardized tests. This creates a structural missingness in the data that is distinct from the standard types of missing data encountered and thus needs to be handled differently. Second, we do not yet have complete compliance data for the multi-child families. For this reason the analyses in this draft are limited to results for

Item Description	% Response
Female guardian's highest level of education	95
Female guardian's ethnicity	94
Female guardian's country of birth	88
Number of years female guardian has lived at current residence	97
Female guardian's employment status	95
Female guardian's religion	94
How often female guardian attends religious services	96
Male guardian's highest level of education	83
Male guardian's ethnicity	81
Male guardian's country of birth	72
Number of years male guardian has lived at current residence	60
Male guardian's employment status	76
Male guardian's religion	71
How often male guardian attends religious services	63
Number of children under 18 living at home	94
Number of children at home attending a public elementary or high school	93
Number of children at home attending a religious private elementary or high school	58
Number of children at home attending a non-religious private elementary or high school	55
Whether there's a daily newspaper in the child's home	90
Whether there's an encyclopedia in the child's home	86
Whether there's a dictionary in the child's home	95
Whether there are more than 50 books in the child's home	92
The main language spoken in the home	92
Whether anyone in the home receives assistance through food stamps	93
Whether anyone in the home receives assistance through welfare (AFDC or public assistance)	89
Whether anyone in the home receives assistance through social security	77
Whether anyone in the home receives assistance through Medicaid	87
Whether anyone in the home receives assistance through Supplemental Security Income (SSI)	79
Yearly income of household before taxes	92

Table 2: Response Rates by Item for Baseline Parent Questionnaire

the 1250 single-child families who applied in grades 1-4.

6 Design

While the lottery used to award scholarships naturally created a randomized design, it also precluded blocking on key variables for added efficiency[4]. Another complication was that evaluation funding only allowed for 1000 treatment families and 1000 control families to be followed up. How to choose the control families from the reservoir of over 4000 families who participated in the lottery but didn't win a scholarship became the focus of the design issues and led to the development of a new experimental design, the Propensity Matched Pairs Design (PMPD). The PMPD is a design which creates matched pairs using the popular propensity score matching technique developed by Rosenbaum and Rubin (1983).

6.1 The Lottery and its Design Implications

The original plan for the lottery included three stages.

1. Interested families would submit applications to the program.

Over 20,000 families participated in the initial application stage. For administrative purposes, applications were batched by the date received into five time periods.

2. All potentially eligible families would be invited to a half-day of screening, which would include confirmation of eligibility, pre-testing of children, and completion of a survey regarding the family's relevant background characteristics.

This plan was followed for the first batch of applicants. However, due to a variety of logistical constraints, coupled with the overwhelming response to the program, not all potentially eligible families were screened in the next four waves. Sampling of applicants had to be performed in order to reduce the number invited to participate in the screening stage. To keep the aggregate probability of receiving a scholarship equal across the time

[4]Randomization within subgroup classifications might have appeared inequitable to the public.

periods, the probability of receiving a scholarship amongst those screened had to be increased to offset the reduced probabilities of being invited to a screening session.

3. Families who completed the screening and whose eligibility was confirmed would be allowed into the final lottery.

Over 5000 families participated in the final lottery. In accordance with the goals of the SCSF, applicants from "bad" schools (schools whose average test scores were below the city-wide median) were given a higher chance of winning a scholarship than those from "good" schools (schools whose average test scores were above the city-wide median). Families from bad schools were to represent 85% of those winning scholarships. This oversampling took place during the lottery for those who applied in the first wave (since there was no sampling performed at the screening stage). In the second through fifth waves, however, the differential selection of those from good or bad schools was largely accomplished in the sampling at the *screening* stage. The implication of this difference is that the treatment and control groups in the last four waves are balanced on the bad/good variable whereas the treatment and initial control groups (i.e., those who did not win a scholarship) from the first wave are unbalanced on the bad/good variable as well as variables correlated with this variable.

6.2 Multi-child Families

The SCSFP was set up so that all eligible siblings of scholarship winners were also offered scholarships. Because of this, families are the unit of randomization, and all matching and subsampling took place at the family level. Since covariate data were collected not only at the family level, but also at the student level, the set of these variables is somewhat different for the families in which more than one child applied to the program ("multi-child" families). That is, since our units of observation are families, yet some data are collected at the student level, multi-child families have more information than single-child families, so the variable "reading test score", for instance, cannot mean the same thing for all families.

For families with more than one child applying, new family variables were created.

These variables were computed across all family members applying. For each family, the average and standard deviation of continuous variables were calculated for initial test scores, age, education expectations and grade level. The mean and standard deviation are based on available values; if only one value is available for a multi-child family, then the standard deviation is missing. For the majority of multi-child families, which are two child families, the original values can be derived from the mean and standard deviation. Binary variables (e.g., bad/good and sex) were recoded as 1 if all responding children in the family responded negatively, 3 if all responding children responded positively, and 2 if responses were mixed. Indicators for the presence of any missing data among all family members for each variable were also created.

6.3 PMPD Versus Randomized Block

The study design provides an opportunity to test empirically the performance of the PMPD. In the first application lottery, in which all apparently eligible applicants were invited to be screened, the ratio of eligible non-winners (control families) to winners (treatment group families) is approximately five to one, an ideal situation for the PMPD. In the second through fifth waves, however, which had smaller control groups due to the limits placed on how many families were invited to be screened, the groups are more nearly equal in size. This latter scenario is more appropriate (given the study design) for a randomized block experiment, with time periods (waves) serving as blocks. Implementing both designs concurrently allows for an empirical comparison of efficiency. However, the PMPD has a more difficult setting in which to achieve balance because of the initial imbalance on the bad/good variable and other baseline covariates correlated with it.

6.4 Design Implementation

The implementation of the two designs proceeded as follows. The data can be conceptualized as being divided into four subgroups based on family size (single vs. multiple

Family Size	Treatment	PMPD	Randomized Block					Total
			2	3	4	5	Subtotal	
Single	Scholarship	404	115	67	82	192	456	860
	Control	2626	72	65	87	135	359	2985
Multi	Scholarship	147	44	27	31	75	177	324
	Control	969	27	23	33	54	137	1106

Table 3: Initial Sample Sizes (unit is a family)

Family Size	PMPD	Rand.Block	Total
Single	353	323	646
Multi	147	177*	354
Overall	500	500	1000

* Only 137 available in control group.

Table 4: Target Sizes for Both Scholarship and Control Samples

children) and design (PMPD vs. randomized block). The initial sample sizes, [5] further broken down by time period, are displayed in Table 3.

The goal was to equalize sample sizes across treatment groups and then, if possible, across blocks, including across single versus multi-child families. It was apparent that we would only be able to approximate this goal in the stratified study. The limiting factor is the number of multi-child control families (137).

Because of financial constraints, we could only follow-up 2000 study participants (a "participant" is a family), and thus some random sub-sampling of lottery winners was performed. Because we had very similar numbers of lottery winners in each design, we targeted a similar number of control families in each design, as seen in Table 4.

[5] These are the sample sizes after removal of 100 families randomly chosen from the control group to receive scholarships for the following academic year, and 100 for the year after that. The additional scholarship offers were used as incentives to increase participation in the follow-up data collection process. New winners were announced following the second and third follow-up testing visits.

6.4.1 Propensity Matched Pairs Design

The strategy for the PMPD was to match 500 sub-sampled scholarship winners from the first time period to 500 controls from the same time period, with separate matching for single and multiple-child families. As a consequence of the dataset being split into two parts (single versus multi-child families), all matching takes place within family size categories. This exact matching on family size produces perfect balance for this variable, which implicitly treats family size as the most important matching variable.

Determinations had been made by the evaluators as to the relative "importance" of the remaining covariates. As described further in Section 6.6.3, importance is judged by a combination of the initial imbalance of a covariate across treatment groups and the perceived strength of the predictive relationship of it to post-randomization outcome measures, which include: the primary outcomes themselves, noncompliance behavior (referring to whether or not a family uses an offered scholarship), attrition from the study, and other types of missing data.

After family size, the most important variable by this definition was judged to be the binary variable for bad versus good school, because it was thought to be highly correlated with the outcomes, and because of the imbalance that occurred in the first time period due to its use in determining lottery winners. It is closely followed in importance by grade level and initial test scores. The remaining covariates are ranked as: ethnicity, mother's education, participation in special education, participation in a gifted and talented program, language spoken at home, welfare receipt, food stamp receipt, mother's employment status, educational expectations, number of siblings (includes children not eligible because of age), and an indicator for whether the mother was foreign born. The final propensity score models, presented in Sections 6.12 and 6.13, were chosen based on the balance created in these variables' distributions across treatment groups. Identification of special variables and the overall ranking of the covariates informed decisions regarding which variables might be appropriate for exact matching, which should receive special treatment in the propensity score method, and what tradeoffs to make in terms of the resulting balance.

The ranking of the variables can be helpful in implementing the propensity score

Family Size	Treatment	PMPD	Randomized Block					Total
			2	3	4	5	Subotal	
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	1000					960	1960

Table 5: Final Sample Sizes

methodology; however, correlations among the variables diminish the importance of the ordering chosen. Therefore the specific ordering chosen may not have a major impact on the creation of matched pairs and should not be viewed as an assumption required for successful implementation.

6.4.2 Sub-Sampling for the Randomized Block Design

We randomly sub-sampled from the cells of the randomized block design to arrive at the final sample sizes, which met the limitation of 1000 families per design. The number sub-sampled were selected to equalize the number of scholarship and control families within blocks, and the number of families across blocks.

1. 133 original single-child lottery winners were randomly withheld for the randomized block design: 43 in time period two, 2 in time period three, 88 in time period five
2. 36 single-child eligible controls were randomly withheld from randomized block design: 5 in time period four, 31 in time period five

The final sample sizes are displayed in Table 5.

6.5 General Propensity Score Methodology

Propensity score matching was introduced by Rosenbaum and Rubin (1983) as a means of creating better balance in observational studies, thereby allowing for valid causal

inference under the assumption of strongly ignorable treatment assignment, i.e., treatment assignment on the basis of the covariates being used to estimate the propensity score. Matching is used as a way of alleviating the biases that can be created by self-selection. As documented in a variety of places (e.g., Rubin 1973, 1979; Roseman 1998), the combination of matching and regression adjustment is typically far superior to either technique alone for controlling bias in observational studies. Not only does matching reduce bias created by the self-selection into treatment groups that occurs in observational studies, it increases efficiency in randomized experiments, such as the one in this study. The extra payoff from matching mostly arises when the linear model underlying regression adjustment is not entirely correct.

The propensity score methods that we use are well-documented and, in the case of no missing data, quite straightforward (Rosenbaum and Rubin 1984). When missing data exist, as they do in this study, extensions of the general methodology (D'Agostino and Rubin 1999) can be implemented. The goal is to balance closely all covariates and patterns of missing data across the treated and matched control groups.

6.6 Complete Data

In the case of complete data, the general strategy is to calculate a "propensity score" for each study participant. This score represents a participant's chance or "propensity" of receiving the treatment (e.g., a scholarship offer),

$$P(Z = 1 \mid X) \ , \tag{1}$$

where Z denotes treatment assignment and X denotes all of the measured covariates (recall, here, fully observed). This probability is straightforward to estimate using logistic regression or linear discriminant techniques.

6.6.1 Matching on the Propensity Score

The propensity scores can be regarded as defining a new covariate value for each individual, which is a function of all of the covariates potentially correlated with the outcomes. In practice the logits of these estimated probabilities are often used because they are

linear in the covariates. Balancing this new covariate generally has the effect of improving the balance of all the other covariates that went into its estimation. A good way to balance propensity scores when the treatment group is much smaller than the control reservoir is to match on propensity scores. Procedurally, this can be accomplished by sorting the treatment group members by their propensity scores and then, one by one, finding for each treated subject, the control group member who has the closest score. Once a match has been made, the chosen control group member is removed from the control reservoir so it cannot be chosen again (Cochran and Rubin 1973). This is called nearest remaining neighbor, or nearest available, matching.

6.6.2 Nearest Available Mahalanobis Matching Within Propensity Score Calipers

The Mahalanobis metric (or distance) between a treatment group member with vector covariate values X_t and a control group member with covariate values X_c (the same set of variables for both), is

$$(X_t - X_c)^T \Sigma^{-1} (X_t - X_c) \ , \tag{2}$$

where Σ is the variance-covariance matrix for these variables, where, in practice, we substitute the pooled sample variance-covariance matrix. A combination of propensity score matching and matching based on the Mahalanobis metric using a subset of variables has many of the advantages of each method (Rubin and Thomas 1996). The combination has been shown to be often superior to either technique used on its own (Rosenbaum and Rubin 1985). With this refinement, as before, propensity scores are calculated for all study participants and then treatment participants are ordered by their propensity scores. Each treatment group member in turn will be initially "matched" to a subset of the control reservoir members whose scores are no more than c propensity score units (e.g., $c = 0.10$ propensity score standard deviations) away from the treatment member's propensity score. Thus the initial matches must fall within a $2c$ length propensity score caliper, symmetric about that treatment group member's score[6]. Mahalanobis matching is used to choose a "nearest neighbor" within this subset of study

[6]This technique is described and illustrated in the context of a real life example in Rosenbaum and Rubin (1985)

participants with respect to several special covariates. The control group member whose values, X_c, of the special covariates minimize the distance from the values, X_t, of the special covariates for the treatment member, is chosen from the subset of controls who fall within the caliper. We include only the continuous covariates most predictive of the outcome variables in the Mahalanobis metric, as discussed in Section 6.6.3.

6.6.3 Special Variables

The more predictive a covariate is of the outcomes of interest, the more crucial is the balance of this covariate across treatment groups. For example, controlling for a covariate (e.g., by balancing) that is uncorrelated with the outcomes plays no useful role, whereas controlling for one that is highly correlated with the outcome will play a crucial role for precise estimation.

Covariates that evaluators are most concerned about balancing receive special treatment in one of two ways. When feasible, exact matches can be required for the most critical of these variables. For instance, if sex were deemed to be the most important variable to balance, when looking at matches for a female treatment group member, no males would be considered. It it is only possible to exact match on discrete variables and only desirable to match on one or two of these. For an example of exact matching in a propensity score context see Rosenbaum and Rubin (1984). Recall that in this study we exact match on family size.

As an alternative to, or in addition to, this exact matching, the Mahalanobis matching within propensity score calipers can be constrained to only a chosen few variables considered more important to balance than the others. Mahalanobis matching is most effective when applied to a small number of essentially continuous covariates (Rosenbaum and Rubin 1985; Gu and Rosenbaum 1993). Matching within propensity score calipers attempts to improve balance for all of the covariates, whereas Mahalanobis matching within calipers attempts to achieve close pair matches on the few special covariates.

6.7 Advantages over ANCOVA (Analysis of Covariance) adjustments

We have already mentioned the benefits of using matching in addition to ANCOVA (regression adjustments) for both bias reduction and precision of estimation. There is another benefit of matching relative to regression adjustment. Adjusting for covariate differences after the experiment has the disadvantage that researchers could settle on the "best" model solely by choosing the one that best supports their a priori biases regarding the issue in question. Matching, on the other hand, uses only covariate balance as a diagnostic; outcomes are not even included in the model, nor are they often even available at the time of matching, as in our application. Therefore, no such researcher bias can occur in the selection of the propensity score model.

6.8 Diagnostics

There are a variety of combinations of the above techniques that will each yield "matched" treatment and control groups. The estimation of the propensity score alone could be accomplished by numerous models, depending on what variables are included and what interactions or non-linear terms are added. Diagnostics, which compare the treatment and control groups with respect to the distributions of the covariates, help the researcher determine which matched control group is superior. Since the goal of the matching is balanced groups, the adequacy of a model or procedure can be judged by treatment versus control group comparisons of sample moments of the joint distribution of the covariates, primarily means and variances, but also correlations. It is often helpful at this stage to have a ranking of covariates in order of perceived importance, beyond just the few selected to be "special" variables. Such a ranking, as described for this study in Section 6.4.1, can help the researcher choose between models with good overall balance that have slight tradeoffs in terms of more or less exceptional balance on specific variables.

6.9 True Versus Estimated Propensity Scores

A surprising fact about the use of propensity scores is that, in general practice, the use of the estimated propensity score typically results in more precise estimates than

the use of the "true" population propensity score. This is especially true when the treatment and control groups are relatively similar initially; the logic is as follows. There are two types of errors that can result from estimates of treatment effect. The first involves systematic biases, which occur when, in expectation, the two groups differ on important characteristics. The second involves conditional biases, which refer to the random differences between groups that average to zero over repeated samples but are nonetheless present in any given sample. Both population and estimated propensity scores effectively reduce the systematic bias in samples; but estimated propensity scores more effectively reduce sample-specific randomly generated bias (Rubin and Thomas 1992). Because a randomized lottery was held to determine scholarship receipt, there is no systematic bias, so estimated propensity scores, in contrast to population propensity scores, more effectively reduce conditional bias.

6.10 Incomplete Data

Techniques to estimate propensity scores in the presence of missing data have been proposed by D'Agostino and Rubin (1999). The type of strategy that is optimal depends upon how the missing data were generated and the relationship of this missingness to the outcomes of interest.

The SCSFP study starts from the advantageous position of a randomized design, within which incomplete baseline data is less problematic than in the case of an observational study. The goal is simply to get the best possible balance on all covariates that we expect to be predictive of outcomes. To the extent that the "missingness" of our covariates is predictive of outcomes, we want propensity score models that include information about the missing data mechanisms (e.g., indicators for the missingness of a particular variable) in order to balance the missingness across treatment groups better than it would be balanced by chance alone. If we believe that this missingness is predictive of the outcomes, then this balance has efficiency implications for our inferences about treatment effects, just as better balance on any other covariate improves efficiency of estimation. In addition, missingness will be used to model compliance status.

As an example, in the SCSFP there were single mothers in the study who refused

to fill out the part of the application survey pertaining to the father of the child. The missingness of these variables could be viewed as a proxy measure for the strength of the relationships in the family and so was hypothesized a priori to be predictive of the outcomes. Therefore this missingness "variable" was included in our propensity model so that we could try to improve its balance across treatment groups.

The other missingness indicator chosen by evaluators as important in this study was that corresponding to mother's education. Investigators think that a missing response to this question reflects a mother's attitude towards education, which could be predictive of educational outcomes, compliance behavior, or subsequent missing data.

The techniques appropriate for including missing data mechanisms in a model are more complicated than those we discussed in Section 6.6. We used a computer program written by Neal Thomas to implement the technique developed by D'Agostino and Rubin (1999), which relies on the ECM algorithm (Meng and Rubin 1993) to calculate propensity scores for each subject, including those with missing covariate values. The ECM algorithm is a variant of the standard EM algorithm which is used in situations where the maximization step is computationally awkward. It replaces the M-step with two or more conditional maximization (CM) steps, each of which has a straight-forward solution.[7].

The Mahalanobis matching within propensity score calipers in the SCSFP project was modified for missing covariate values as follows. If possible, for the matched control, the same missing pattern was required. If no such matched control was found, we exact matched on the design variable bad/good school, which was fully observed. If a matched control still was not found, we would have matched on the propensity score alone; however, this situation never occurred.

[7]For the general location model (often used with missing data e.g., Little and Rubin (1987) and Schafer (1997)), one CM-step gets maximum likelihood estimates for the parameters in the normal distributions conditional on the parameters for the log-linear model (cell probabilities for the contingency table) and a second CM-step obtains estimates for the log-linear model conditional on the parameters of all of the multivariate normal distributions. More CM-steps are often used within the log-linear model portion to avoid running the Iterative Proportional Fitting (IPF) to convergence at each iteration of the ECM algorithm. Bishop, Fienberg, and Holland (1975).

6.11 Relative Strengths of Designs – Diagnostics

We can judge the relative strengths of our designs through diagnostics that measure balance in various ways. Results from the PMPD are contrasted with results from both the randomized block design (2nd through 5th time periods) and a stratified random sample chosen from the control reservoir in the first time period. The stratified random sample was randomized within bad/good school categories; 85% of the children were chosen to be from bad schools and 15% from good schools. This comparison was chosen because it represents the most likely alternative to the PMPD design that MPR would have implemented.

6.12 Single Child Families

Following the criteria discussed in Section 6.4.1, a model for the propensity score was chosen. The contingency table for the categorical variables ethnicity (Hispanic/Black/other), religion (Catholic/other), participation in gifted program, participation in special education, and winning a scholarship, is constrained by a log-linear model that allows for two-way interactions. The continuous portion of the general location model places an additive model across contingency table cells on the means of the following variables: language (spanish/english), whether or not father's work status is missing, participation in food stamp program, participation in Aid to Families with Dependent Children (AFDC), bad/good school, mother's birth location (U.S./Puerto Rico/other), sex, number of eligible children in household, income, mother's education, math scores and grade level. Mahalanobis matching was done in 0.10 calipers of (linear) propensity score standard deviations on the two test score variables and the grade level variable; the bad/good variable also played a special role in the Mahalanobis matching as described in Section 6.10. For algorithmic efficiency, indicator variables for discrete variables that are fully observed (such as bad/good), and any of their interactions, can be treated as continuous with no loss of generality. This is preferable as it reduces the effective dimensionality of the model.

The resulting balance for variables designated by the evaluation team to be most pre-

dictive of outcomes[8] is given in Table 6. In the table, "z-stat" stands for the z-statistic corresponding to the difference in means between the two groups for a covariate[9]. The results for the PMPD are compared to the results for the randomized block design and to the results for stratified random sample (stratified on bad/good) of the same size from the pool of all potential matching subjects.

Overall, the resulting balance from the PMPD is quite good. Compared to the randomized block design, the PMPD has lower absolute z-scores for 16 variables, higher z-scores for only 4. It is beaten by the simple random sample for 6 variables; however, the gains when PMPD beats the simple random sample are generally larger than the gains when the simple random sample beats PMPD.

Propensity score theory predicts a gain in efficiency for differences in covariate means over simple random sampling by a factor of approximately two (Rubin and Thomas 1992, 1996). We have constructed half-normal plots of the Z-statistics displayed in Table 6 which were standardized by the usual two-sample variance estimate, which assumes random allocation to treatment groups. Therefore, we expect these Z-statistics to follow the standard normal distribution when the assumptions of random allocation are true (thus the Z-statistics are expected to fall on the solid line with slope 1 in each diagram). If the observations fall above the line with slope 1, they originate from a distribution with *larger* variance than we are using to standardize the differences, because they are systematically more dispersed than the corresponding quantiles of the standard normal. If they fall below that line, they originate from a distribution with *smaller* variance than we are using to standardize the differences because they are systematically less dispersed than the the standard normal.

For Figure 1, the dotted line has slope $1/\sqrt{2}$, corresponding to the normal distribution with variance 1/2. This figure thus reveals that the gains predicted by Rubin and Thomas (1992) are fairly closely achieved for the study of single-child families. In

[8]The list of all variables included in the final analysis is displayed in Table 16 in the Appendix.

[9]This is calculated for each covariate, x, as

$$\frac{\bar{x}_t - \bar{x}_c}{\sqrt{\hat{\sigma}_t^2/n_t + \hat{\sigma}_c^2/n_c}}$$

where t and c subscripts denote sample quantities from the treatment and control groups, respectively.

Variable	Stratified Random Sample	PMPD	Randomized Block
bad/good	-0.98	0.11	0.21
grade level	-1.63	-0.03	-0.39
reading score	-0.38	0.48	-1.05
math score	-0.51	0.20	-1.37
ethnicity	1.80	1.59	1.74
mom's education	0.16	0.09	1.67
special education	0.31	-0.17	0.22
gifted program	0.42	-0.13	0.75
language	-1.06	-1.03	-0.44
afdc	-0.28	0.83	-1.57
food stamps	-1.08	0.94	-1.31
mother works	-1.26	-1.18	0.40
educ. expectations	0.50	0.57	0.19
children in household	-1.01	0.41	-1.02
birth location	0.49	-1.40	-0.69
length of residence	0.42	0.66	-0.78
dad's work missing	1.09	0.00	0.16
religion	-1.84	-0.74	-0.80
sex	0.88	0.76	0.53
income	-0.38	0.74	-1.21
age as of 4/97	-1.57	-0.47	-0.87

Table 6: Balance: Single-Child Families

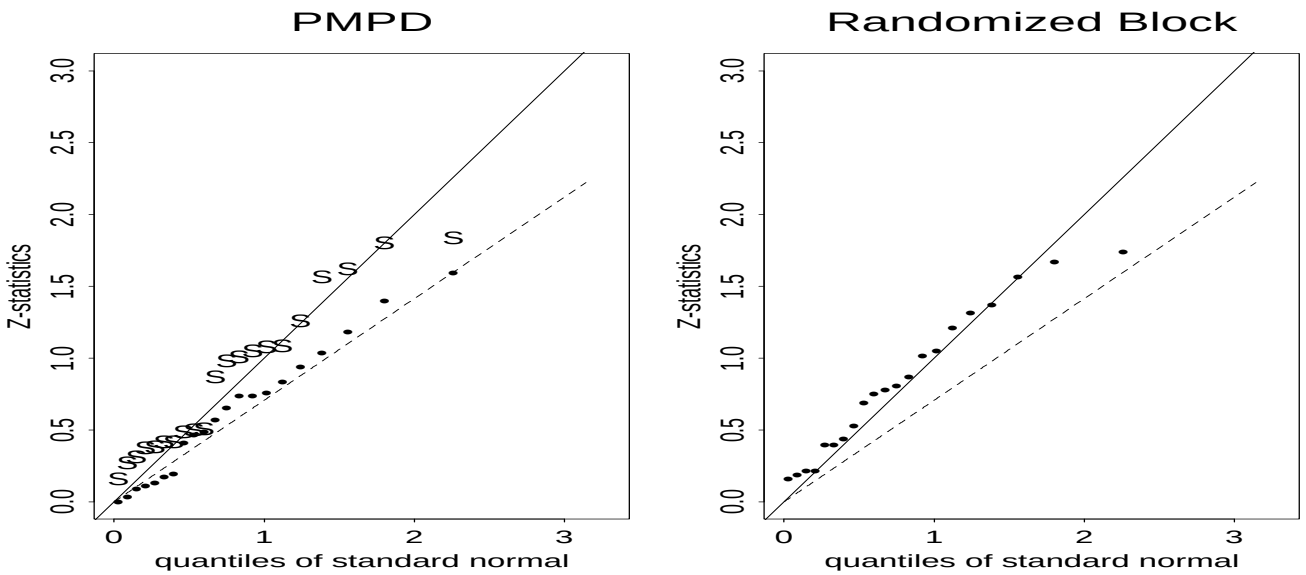

Figure 1: Half-Normal Plots of Z-Statistics for Single-Child Families

fact, compared to the stratified random sample (shown by the "S" points), we see even greater gains. These results can be contrasted with those from the randomized block experiment, which are consistent with the theoretically predicted slope of 1 given by the the solid reference line.

Since the variance in the difference in means is reduced by a factor of two, this is equivalent to increasing the sample size by a factor of two for these variables. Although this holds for any linear combination of the measured covariates, in practice outcome variables are not perfectly predicted by these variables, resulting in a less dramatic improvement in efficiency (Rubin and Thomas 1996).

6.13 Multi-Child Families

Following the criteria discussed in Section 6.4.1, a propensity model was chosen. The contingency table for the categorical variables (ethnicity, religion, sex, birth location, and winning a scholarship) is constrained by a log-linear model that allows for two-way interactions. The continuous portion of the general location model places an additive model across contingency table cells on the means of the following variables:

participation in gifted program, participation in special education, language, whether father's work status is missing, participation in food stamp program, participation in AFDC, bad/good, number of eligible children in household, income, mother's education, mother's length of residence, mother's work status, average and standard deviation of children's ages, average and standard deviation of educational expectations, average and standard deviation of math and reading scores, and average and standard deviation of grade. Mahalanobis matching was done in 0.10 calipers of linear propensity score standard deviations on the four test score variables and the two grade level variables; the bad/good variable also played a special role in the Mahalanobis matching as described in Section 6.10.

The resulting balance of the design as compared with the corresponding randomized block design, and a stratified random sample of the potential matches is displayed in Table 7. The initial imbalance in the bad/good variable is also present with the multi-child families, but the PMPD still achieves very good overall balance. Compared to the randomized block design, the PMPD has lower absolute z-scores for 18 variables, higher z-scores for 8. The win/loss ratio is 17 to 9 for the comparison with the stratified random sample, although, again the gains when PMPD beats the stratified random sample are generally larger than when stratified random sample beats PMPD.

Half-normal quantile-quantile plots for the multi-child families in both experiments, displayed in Figure 2, are similar to those for single-child families. Gains in efficiency by a factor of two appear to be achieved by the PMPD over the randomized block design and by slightly greater than two over the stratified random sample because of the initial imbalance in the bad/good variable.

Although the special test score variables are not quite as well balanced in the PMPD as in the randomized block design for the multi-child families (probably due to correlations between these and the bad/good variable), they are still well balanced. Furthermore, the high correlation commonly seen between pre- and post-test scores makes this variable a prime candidate for covariance adjustments within a linear model to take care of the remaining differences between groups. For the single-child families, the PMPD is clearly superior in terms of test score variable balance.

Variable	Stratified Random Sample	PMPD	Randomized Block
bad/good	-3.81	-0.98	0.15
avg. grade level	-0.27	0.38	0.23
s.d. grade level	-0.19	-0.40	0.58
avg. reading score	-1.06	0.91	-0.23
s.d. reading score	-0.90	1.23	-2.20
avg. math score	-0.56	0.82	0.32
s.d. math score	-1.02	0.33	-1.11
ethnicity	-1.03	0.20	2.09
mom's education	-0.27	-0.21	-0.22
special education	-0.67	-0.11	0.68
gifted program	-0.85	-0.07	-0.52
language	1.13	0.92	-0.64
afdc	-1.24	0.13	3.42
avg. age	-0.38	0.48	0.66
s.d. age	0.09	0.00	0.38
avg. educ. exp.	-0.81	0.49	-0.71
s.d. educ. exp.	-1.59	-0.10	0.94
children in household	0.39	-0.40	-0.13
income	0.93	0.13	2.01
religion	0.01	-0.97	-0.66
length of residence	-1.29	0.54	1.31
dad's work missing	0.39	0.70	1.73
food stamps	-2.06	-0.35	2.58
mom works	1.29	0.73	-0.49
birth	0.20	-0.42	1.34
sex	-0.84	-0.17	-1.43

Table 7: Difference in Means Z-Statistics: Multi-Child Families

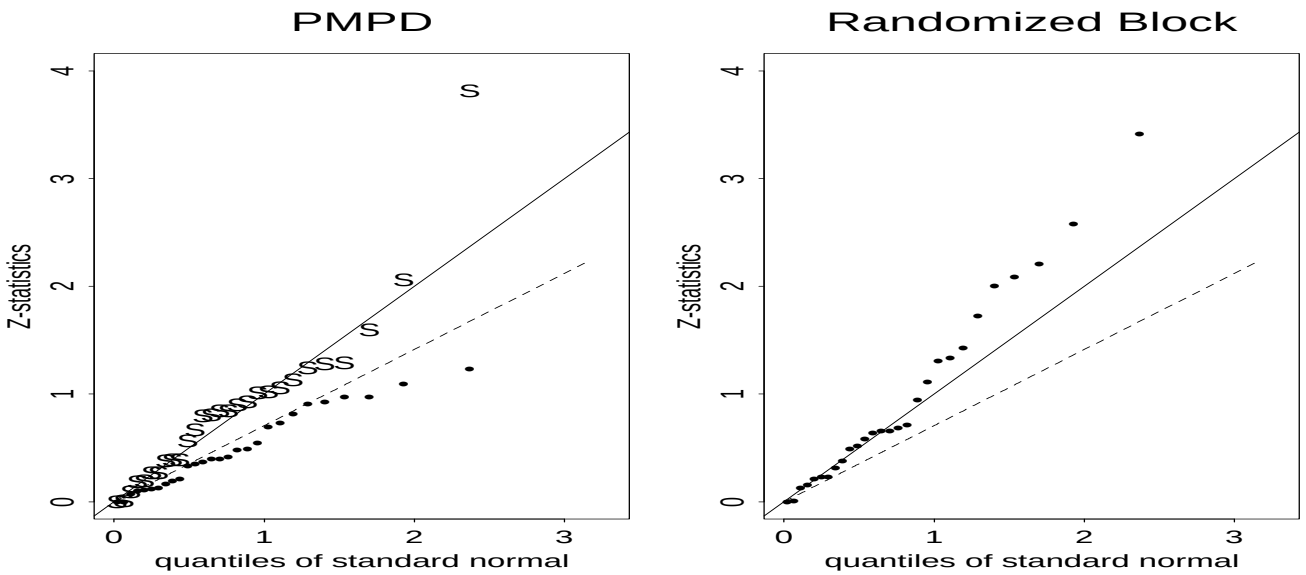

Figure 2: Half-Normal Plots of Z-Statistics for Multi-Child Families

It is worthwhile to note that all of the calculations in the section were performed on an available-case basis to provide statistics comparing balance. They are not directly relevant for drawing causal inference.

7 Broken Randomized Experiments

It is important to realize that our randomized experiment does not really randomize the treatment of public and private schools but rather it randomizes the "encouragement" to attend a private rather than a public school by offering to provide some financial support ($1400) to do so. In some encouragement studies interest may focus on the effect of encouragement itself, but more often when randomized encouragement designs are used, interest focuses on estimating the effect of the treatment being encouraged, here, attending private versus public schools (or participation in the scholarship program). If there were perfect compliance, so that all those encouraged to get the new treatment got it, and all those who were not so encouraged received the standard treatment, then the effect being estimated would be typically attributed to whatever was viewed as the "active" ingredient in the treatment condition. But encouragement designs do not

anticipate anything approaching full compliance, and so there is the opportunity to try to estimate different effects for encouragement and the active treatment.

In recent years, there has been substantial progress in the analysis of encouragement designs, based on building bridges between statistical and economic 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 1978, 1990) . In economics, the technique of "instrumental variables" (IV) due to Haavelmo (1943, 1944) was a main tool of causal inference in the type of non-randomized studies that dominate economics. Angrist, Imbens, and Rubin (AIR, 1996) showed how the approaches were completely compatible, thereby clarifying and strengthening each. The result was the interpretation of the IV technology as a way to attack a randomized experiment with noncompliance, such as a randomized encouragement design.

Imbens and Rubin (1997a) showed how the Bayesian approach to causal inference in Rubin (1978) could be extended to handle simple randomized experiments with non-compliance, and Hirano, Imbens, Rubin, and Zhou (1999) showed how the approach could be extended to handle fully observed covariates, and applied it to an encouragement design in which doctors were randomly encouraged to give flu shots to at-risk patients.

Our setting is far more complex, because we have missing covariates and multivariate outcomes that are sometimes missing as well. The basic structure for our type of problem was outlined in Barnard, Du, Hill, and Rubin (1998), but our situation is more complex than that because we have a more complicated form of noncompliance – some children attend private school without receiving the monetary encouragement; it is slightly less complicated because we currently have outcomes from only one post-treatment time point. As in Frangakis and Rubin (1999) and Barnard *et al.* (1998), because of the problems described in Section 5.1 we need to make some assumptions

about the missing data process and treatment effects for the non-compliers.

The first assumption we make has been called "compound exclusion" by Frangakis and Rubin (1999), when they generalized the exclusion restriction in economics. The way AIR define the exclusion restriction is as follows: for those subjects whose behavior cannot be changed by the random assignment in this experiment (i.e., the encouragement to attend private schools), their outcome scores are unaffected by the assignment. That is, for those whose behavior is unaffected by assignment, their outcomes are also unaffected. Thus, under this assumption, the always takers, those who, in the context of this experiment, will attend private school whether or not encouraged to do so, will have the same outcomes (test grades) in the private school they are attending whether or not they were encouraged. Analogously, those who, in the context of this experiment, will not attend private schools whether or not they are encouraged to do so, will have the same test grades whether or not they are encouraged to attend private school. Actually, this is what Imbens and Rubin (1997a) call "weak exclusion" because it says nothing about the compliers in this experiment, whereas the strong exclusion restriction, which is the traditional economic version, adds the assumption that differences in outcomes for assigned and not assigned compliers is due to treatment exposure and *not* assignment to be encouraged or not. The compound exclusion restriction of Frangakis and Rubin (1999) extends the weak exclusion restriction to apply to the missing data pattern of the outcomes as well as the values of the outcomes.

The exclusion restriction focuses attention on the "complier average causal effect" (CACE), which is the average causal effect of assignment for the compliers, rather than the more traditional "intention to treat" effect (ITT), which is the average casual effect of assignment for all subjects. Under exclusion, the average causal effects of assignment for never takers and always takers is zero, so if this assumption is correct, the ITT effect is the weighted average of the CACE and zero.

The second assumption we make has been termed "latent ignorability" of the missing data mechanism by Frangakis and Rubin (1999). Ignorability of the missing data mechanism (Rubin, 1976, Little and Rubin, 1987) basically means that the missingness of the data, given the observed values, is not dependent on missing values themselves

or the parameters of the data distribution. Latent ignorability states that ignorability holds if a latent variable were fully observed, here the true compliance status of each subject (complier, never taker, always taker). Notice that we have implicitly made another assumption, namely that there are no defiers, subjects who when encouraged to attend private school will not, but when not encouraged to do so will.

As Imbens and Rubin (1997a) and Hirano, Imbens, Rubin, and Zhou (1999) show, none of these assumptions are needed for a valid Bayesian analysis when faced with noncompliance, but they can dramatically simplify the analysis and sharpen posterior inferences. In fact, this is one of the dramatic advantages of the Bayesian approach to this problem: the issue of "identifiability" is put in its proper perspective. It is largely irrelevant to inference if the likelihood function has one mode rather than a small ridge – the important inferential issue is the size of a, e.g., 90% interval, and not whether or not an ϵ% interval is unique for all positive ϵ.

A final point about our situation, with noncompliance to encouragement and missing outcomes, is that even if the focus of estimation is on the ITT effect and not CACE, one cannot use ad hoc methods to estimate the ITT effect without incurring bias. Under compound exclusion and latent ignorability, Frangakis and Rubin (1999) show that an estimator analogous to the IV estimator can be used to estimate the CACE and thereby the ITT effect essentially without bias. Of course, our Bayesian analysis does this automatically.

8 Complete Case Analyses

To provide a first look at the data using types of analyses that are commonly performed, we present results from complete-case analyses. Specifically, we perform simple analyses only on the 752 individuals from single-child families (this represents 56% of all single-child families) who had the following variables fully observed: reading post-test score[10], math post-test score, reading pre-test scores and math pre-test scores (this automatically excludes children who applied in kindergarten). Covariates that were fully

[10] All test scores are normal curve equivalents of national percentile rankings within grade.

observed for all study participants were also included in the analyses: application wave, grade level, and quality of school (bad/good).

We present these results before we discuss the formal notation because, although we do not regard a complete-case approach as inferentially valid in any generality, it is straightforward to explain. In addition, complete-case analyses are standard practice in many disciplines, have certain advantages in some settings (see, for instance, Little and Rubin 1987), and will serve as a useful comparison to our more fully Bayesian analysis. For conciseness of exposition, we use the common "significant" and "insignificant" terminology to summarize the evidence for 95% interval estimates including zero.

8.1 ITT Analysis with Complete Cases

The ITT analysis on complete cases (n=752) for the variables listed in Table 8 demonstrates significant positive effects of scholarship assignment (winning the lottery) on reading scores for students who applied in 4th grade while in bad schools. The insignificant remaining treatment interactions indicate that this is most likely a generally significant effect across bad/good and grade combinations. Math scores demonstrate no such significant treatment effect. Attending a good school prior to the study appears to be related to more positive post-test scores than attendance at a bad school. Not surprisingly, reading and math pretest scores are quite predictive of post-test scores. Grade level, however, seems to have a less consistent relationship with post-test scores. For both reading and math post-test scores we see a significant grade 3 association (relative to grade 4), but there is also a significant positive association for the second grade (relative to 4th grade) for reading scores.

8.2 CACE estimation using OLS and Maximum Likelihood

The easiest and most standard approach to CACE estimation lies within the instrumental variable (IV) framework from the economics literature. A standard estimation approach in this framework in two-stage least squares. Alternately, maximum likelihood estimates for this mixture model can be accomplished, for example, via the EM algorithm. We expect the maximum likelihood estimates to be more trustworthy than

	READING		MATH	
	coef.	s.e.	coef.	s.e.
intercept	11.51	2.53	11.12	2.66
treatment	5.44	2.66	2.73	2.79
application. wave				
1	-1.27	1.62	1.72	1.71
2	-4.00	2.18	-1.89	2.29
3	-2.89	2.26	2.64	2.37
4	-3.27	2.20	-2.06	2.30
good school	4.01	2.29	5.01	2.41
reading pretest	.46	.03	.34	.03
math pretest	.15	.03	.30	.04
grade				
1	4.90	2.43	-2.53	2.55
2	2.71	2.46	-2.18	2.58
3	8.65	2.36	4.96	2.48
treatment×grade				
1	-4.59	3.39	-.17	3.56
2	-6.40	3.39	-3.32	3.56
3	-4.90	3.37	.63	3.54
treatment×good	.42	3.15	-.21	3.30

Table 8: Intention-to-treat Analysis (OLS)

the OLS estimates because it has been shown that in this setting OLS estimation can lead to inconsistencies such as parameters estimates outside the parameter space (see, for instance, Imbens and Rubin, 1997a).

The first set of analyses define the treatment to be the effect of the scholarship program (receiving a scholarship as well as help in identifying an appropriate school). Therefore always takers are defined to not exist in this setting. These results, presented in Table 9, compare children participating in the program and those not participating (some of whom, however, may be attending private school) all of whom would participate in the program if given the chance.

The most consistently significant predictors of post-test scores across analyses are reading and math pre-test scores. Both methods estimate significant positive impact of the program on reading scores for students applying from bad schools in fourth grade (and third-grade reading scores for OLS). Relative to these we see added effects for those in third grade according to the OLS estimates. Effects on the math scores appear to be swamped by noise and no significant differences from the baseline estimate for the remaining interactions. Good schools appear to be correlated with higher post-test scores according to the OLS analyses but this relationship is not strong enough for statistical significance; large standard errors make this relationship difficult to assess in the ML analysis. Grade level has inconsistent associations with post-test scores across the analyses with generally stronger and more positive results on reading test scores, with significance for the third grade relative to the fourth. In general, the ML estimates have larger standard errors than the OLS estimates; in fact, there are no substantive significant estimates for the math scores in the ML analysis. This is due to the fact that the ML model we fit was less constrained than the OLS models[11].

The second set of analyses define the treatment to be private school attendance, therefore this analysis includes compliers, never takers and always takers; results are shown in Table 10. The results for the OLS analyses are almost identical to those from

[11]Each compliance group was allowed a distinct set of regression parameters. In addition, standard errors produced by such OLS models have been criticized for being anti-conservative in some settings (see Imbens and Rubin 1997b). The standard errors we present have been "Huber-corrected" (Over *et al.* 1995) and therefore may be more robust than otherwise.

	READING				MATH			
	OLS-IV[1]		ML-IV[2]		OLS-IV[1]		ML-IV[2]	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef	s.e.
intercept	11.44	2.59	8.86	3.98	11.09	2.79	8.32	4.63
treatment	5.96	2.86	8.92	3.58	2.99	2.96	3.25	4.73
application wave								
1	-1.39	1.58	-.06	3.22	1.66	1.77	4.05	3.08
2	-4.16	2.38	2.12	3.71	-1.97	2.22	-.31	3.54
3	-2.71	1.98	-2.24	3.93	2.73	2.26	-.93	4.41
4	-3.03	2.00	-4.48	3.62	-1.94	2.26	-1.53	4.06
good school	3.97	2.25	4.03	4.41	4.99	2.80	-4.40	6.29
reading pretest	.47	.04	.46	.06	.34	.04	.34	.07
math pretest	.14	.03	.13	.07	.30	.04	.31	.07
grade								
1	4.88	2.62	12.13	4.18	-2.54	2.64	.60	4.97
2	2.75	2.37	3.12	3.60	-2.16	2.73	-8.13	6.43
3	8.66	2.41	6.87	3.88	4.97	2.75	3.82	5.50
treatment×grade								
1	-3.88	3.28	-13.94	5.50	.19	3.19	-3.74	6.22
2	-6.05	3.10	-7.62	4.53	-3.14	3.34	3.26	7.02
3	-4.45	3.12	-3.28	4.93	.85	3.37	3.30	6.47
treatment×good	.66	3.11	-1.79	5.86	-.09	3.50	7.39	6.99

1 Calculated using Stata (StataCorp. 1997).

2 Calculated using Mplus (Muthen and Muthen 1998).

Table 9: Effect of Scholarship Program – analysis with only never takers

	READING				MATH			
	OLS-IV		ML-IV		OLS-IV		ML-IV	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef	s.e.
intercept	11.55	2.77	29.74	5.26	10.71	3.01	16.67	5.20
treatment	6.78	3.20	6.00	4.43	3.41	3.35	4.33	4.43
application wave								
1	-1.44	1.59	-7.69	3.81	1.63	1.78	-.96	3.70
2	-3.90	2.37	-12.52	5.09	-1.84	2.21	-3.12	3.95
3	-2.49	2.01	-5.08	4.49	2.84	2.28	2.20	4.52
4	-2.91	2.28	-15.01	4.72	-1.88	2.27	-4.45	4.82
good school	4.26	2.28	8.13	3.75	5.14	2.81	10.58	6.43
reading pretest	.47	.04	.07	.10	.34	.04	.07	.10
math pretest	.14	.03	-.18	.12	.30	.04	.17	.17
grade								
2	4.99	2.60	3.42	4.78	-2.49	2.65	2.13	5.36
3	2.77	2.32	-9.04	5.42	-2.15	2.72	-5.87	5.21
4	8.68	2.39	11.92	4.41	4.96	2.74	-4.81	5.61
treatment×grade								
2	-3.92	3.26	-11.04	6.39	.17	3.19	-6.82	6.37
3	-6.03	3.06	24.11	7.91	-3.14	3.32	13.17	10.61
4	-4.39	3.09	-2.02	7.29	.89	3.35	24.70	8.31
treatment×good	.42	3.13	9.90	5.63	-.21	3.52	3.33	7.98

Table 10: Effect of Private School – analysis with never takers and always takers

the previous analyses. This is not surprising given that always takers account for only approximately 9% of the study population. The ML point estimates are generally quite similar to those for the previous analysis however the standard errors are much larger due to the addition of the full set of always-taker-specific parameters.

9 Overall Strategy

An ideal scenario for obtaining valid causal inferences for a binary treatment is the following: (1) the data arise from a randomized experiment with two treatments; (2) the outcome variables are fully observed; (3) there is full compliance with the assigned treatment; and (4) the blocking variables are fully observed. ; and (5) the background variables are fully observed. Aspect (5) is useful for doing covariate adjustment and subpopulation analyses. For this ideal scenario, there are standard and relatively simple methods for obtaining valid causal inferences. In reality, however, this scenario rarely occurs. Clearly, it does not occur in the SCSFP.

Deviations from the ideal scenario that occur frequently and are present in the SCSFP are the following: (6) there exist missing values in the outcomes; (7) there exist missing values in the background variables; and (8) there is noncompliance with assigned treatment. The standard methods for analyzing the ideal scenario of (1)–(5) generally fail when aspects (6)-(8) are present. Handling these additional complications in a valid and general manner is difficult and beyond the current state of the art. Here we present an extremely general data template allowing (6)-(8). When the observed data can be made to conform to this template, we are able to obtain valid causal inferences by creating multiple imputations and by using complete-data tools designed for the standard scenario described in (1)–(5). Our current model will return us to the scenario consisting of (1)–(4).

10 Pattern Mixture approach

Suppose that we have policy relevant covariates that are fully observed, in addition to other covariates, that may be very important for precision of estimation, which

are only partially observed. Within the context of a randomized experiment, we can conceive of a sub-experiment within each pattern of missing data, which is also perfectly randomized (just as when we divide a completely randomized experiment into males and females, for instance). That is, pre-treatment missing data patterns can be considered covariates themselves. Consequently, an attractive practical alternative when dealing with missing covariates that are not policy relevant in the above sense is to adopt a pattern mixture approach to the analysis. Of course if a policy relevant covariate is missing then this approach is not satisfactory and that covariate must become part of the model. Fortunately in our setting the major policy relevant covariates (grade, quality of school), on which decisions regarding viability of new programs may be made, are fully observed. The covariates that are important but missing are individual-level characteristics such as pre-test scores which we do not consider policy-relevant in the above sense because they're no new program would be predicated upon these (e.g., it's difficult to imagine that pre-test scores would used as eligibility criteria for a program whereas school quality is used all the time).

In principle of course our Bayesian models can include all covariates that are partially missing, but the analysis becomes even more complicated.

11 Notation for our Data Template

In our template, we assume that for the i^{th} subject, where $i = 1, \ldots, n$, we have the following random variables:

1. Binary indicator of treatment assignment

$$Z_i = \begin{cases} 1 & \text{if subject } i \text{ is assigned to treatment group,} \\ 0 & \text{if subject } i \text{ is assigned to control group.} \end{cases}$$

   Z is the n component vector with i^{th} element Z_i.

2. Binary indicator of treatment receipt

$$D_i = \begin{cases} 1 & \text{if subject } i \text{ received treatment,} \\ 0 & \text{if subject } i \text{ received control.} \end{cases}$$

Because D_i is a post-treatment-assignment variable, it has a potential outcome formulation, $D_i(Z_i)$, where $D_i(0)$ and $D_i(1)$, respectively, refer to the values when assigned control and when assigned treatment.

3. Compliance status

$$C_i = \begin{cases} c & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 1 \\ n & \text{if } D_i(0) = 0 \text{ and } D_i(1) = 0 \\ a & \text{if } D_i(0) = 1 \text{ and } D_i(1) = 1 \end{cases}$$

$C_i = c$ denotes a "complier," a person who will take the treatment if so assigned and will take control if so assigned. $C_i = n$ denotes a "never takers," a person who will not take the treatment no matter the assignment. $C_i = a$ denotes an "always takers," a person who will always take the treatment no matter what the assignment. This template rules out the possibility of "defiers," those who will always do the opposite of what they are assigned, i.e. those i for whom $D_i(0) = 1$ and $D_i(1) = 0$. C denotes the n component vector with i^{th} element C_i.

4. $2P$-component vector of potential outcomes[12], Y_i^{po},, which is composed of two P-length vectors, $Y_i(0)$ and $Y_i(1)$, where

$$\begin{aligned} Y_i(0) &= (Y_{i1}(0), \ldots, Y_{iP}(0)), \text{ and} \\ Y_i(1) &= (Y_{i1}(1), \ldots, Y_{iP}(1)). \end{aligned}$$

Here $Y_{ip}(0)$ is the p^{th} outcome variable corresponding to assignment to the control group for the i^{th} subject; $Y_{ip}(1)$ is the p^{th} outcome variable corresponding to assignment to the treatment group for the i^{th} subject. In other words, for each subject there are P outcome variables, Y_1 through Y_P, and has two potential values: one corresponding to each of the treatment assignments. $Y(0)$ and $Y(1)$ are used to denote the $n \times p$ matrices of potential outcomes corresponding to control and treatment assignment respectively.

[12] In general, out template allows for repeated measurements over time. However, currently we have data from one pre-treatment time point and only one post-treatment time point and our notation reflects this simplification.

At times we will refer simply to the P-component vector of outcomes that we *intend* to observe for a person, i.e.,

$$Y_i^{\text{int}} = Y_i(Z_i).$$

For convenience, we will henceforth refer to Y_i^{int} as simply Y_i with corresponding elements

$$Y_i = (Y_{i1}, \ldots, Y_{iP})$$

In addition, Y represents the n by P matrix of intended outcomes for all study participants. $Y_{\cdot p}$ is the p^{th} column in this matrix.

5. $2P$-component vector of response patterns for potential outcomes.

$$R_{y_i}(t) = \begin{cases} 1 & \text{if } Z_i = t \text{ and } Y_{ip}(t) \text{ is observed,} \\ & \text{or, if } Z_i \neq t \text{ but } Y_{ip}(t) \text{ would be observed if } Z_i = t, \\ 0 & \text{if } Z_i = t \text{ and } Y_{ip} \text{ is not observed,} \\ & \text{or, if } Z_i \neq t \text{ but } Y_{ip}(t) \text{ would be unobserved if } Z_i = t, \end{cases}$$

These indicators are themselves potential outcomes because we can only observe response indicator $R_{y_i}(t)$ for individual i if $Z_i = t$.

6. P-component outcome response pattern associated with each Y_i

$$R_{y_i} = (R_{y_{i1}}, \ldots, R_{y_{iP}}),$$

where

$$R_{y_{ip}} = \begin{cases} 1 & \text{if } Y_{ip} \text{ is observed,} \\ 0 & \text{if } Y_{ip} \text{ is not observed.} \end{cases}$$

R_{y_i} indicates which of the P outcomes are observed and which are missing for subject i. R_y denotes the $n \times P$ matrix of missing outcome indicators for all study participants. $R_{y \cdot p}$ is the p^{th} column in this matrix.

7. K-component vector of fully observed background and design variables

$$W_i = (W_{i1}, \ldots, W_{iK}),$$

where W_{ik} is the value of fully observed covariate k for subject i. W is the $n \times K$ matrix of fully observed covariates. $W_{\cdot k}$ is the k^{th} column in this matrix. In this study, application wave, the quality of the school the child attended at time of application (bad/good), and grade level are fully observed.

8. Q-component vector of partially observed background and design variables

$$X_i = (X_{i1}, \ldots, X_{iQ}),$$

where X_{iq} is the value of covariate q for subject i. X represents the n by Q matrix of covariates for all study participants. $X_{\cdot q}$ is the q^{th} column in this matrix. In addition, $X^{(\text{cat})}$ refers to the subset of covariates that are categorical and $X^{(\text{cont})}$ refers to the subset of covariates that are continuous.

9. Covariate response pattern associated with X_i

$$R_{x_i} = (Rx_{i1}, \ldots, Rx_{iQ}),$$

where

$$R_{x_i} = \begin{cases} 1 & \text{if } X_{iq} \text{ is observed,} \\ 0 & \text{if } X_{iq} \text{ is not observed.} \end{cases}$$

R_{x_i} indicates which covariates are observed and which covariates are missing out of the Q possible covariates for subject i. R_x denotes the $n \times Q$ matrix of covariate missing data indicators for all study participants. $R_{x \cdot q}$ is the q^{th} column in this matrix.

This observed data template is extremely general, allowing arbitrary response patterns for the outcomes and covariates.

12 The Model

Our full model needs to simultaneously (1) represent a reasonable approximation to the sampling distribution of the (complete) data, (2) be comprehensive enough to justify

our assumptions about the missing data process, (3) incorporate the constraints imposed by the randomization, (4) incorporate the constraints imposed by the exclusion restriction, and (5) incorporate the conditional independence structures imposed by the latent ignorability.

We describe the model first by giving its structural assumptions, that is assumptions that can be expressed without reference to a particular distributional family. Then we describe the assumptions of the particular parametric model we assume.

12.1 Structural Assumptions

We now formalize the structural assumptions of our model (some of which were previously introduced in Section 7) and discuss their plausibility for this study.

12.1.1 SUTVA

A standard assumption made in causal analyses of this kind is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978, 1980, 1990). This assumption implies that one unit's treatment assignment does not affect another unit's outcomes and there are no versions of treatments. Formally, SUTVA is satisfied if $Y_i(Z) = Y_i(Z')$ and $D_i(Z) = D_i(Z')$ if $Z_i = Z'_i$, where Z' is the n-length vector with i^{th} element Z'_i. In this study, for SUTVA to be violated, the fact that one family won a scholarship or didn't would have to affect outcomes such as another family's choice to attend private school or their children's test scores. It does not seem a terribly strong assumption to disallow such effects.

12.1.2 Randomization

We assume scholarships have been randomly assigned. This implies

$$p(Z \mid Y(1), Y(0), X, W, C, R_y(0), R_y(1), R_x, \theta) = p(Z \mid \theta) = p(Z),$$

where we drop the dependence on θ because there are no unknown parameters governing the treatment assignment mechanism. This "assumption" should be trivially

satisfied given that MPR administered a lottery to assign scholarships to families and the differential sampling weights for school type and application wave are known.

12.1.3 Missing data process assumption – Latent Ignorability

We assume that missingness can be predicted by observed covariates within subgroups defined by compliance status. This is defined as "latent ignorability" of the missing data mechanism, formally,

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

where X^{obs} comprises the elements of the covariate data matrix X that are observed and θ is a generic parameter representing any model. Note that this is a *non-ignorable* missing data mechanism.

Recall that latent ignorability differs from standard ignorability (Rubin 1978; Little and Rubin 1987) because it conditions on something that is (at least partially) unobserved or latent, in this case, compliance status, C. This is a more reasonable assumption than standard ignorability because it seems quite likely that the groups of people defined by compliance status would behave differently with regard to whether or not they fill out surveys or show up for post-tests.

12.1.4 Noncompliance process assumption I – Compound Exclusion

In order to discriminate among compliers, never takers, and always takers, we need to make an assumption about their behavior. Given that never takers and always takers will participate in the same treatment (control or treatment, respectively) regardless of what they were randomly assigned, it seems plausible to assume that their outcomes and missing data patterns will not be affected by treatment assignment. The compound exclusion restriction reflects this assumption, formally, as

$$p(Y(1), R_y(1) \mid X, R_x, W, C = n) = p(Y(0), R_y(0) \mid X, R_x, W, C = n),$$

for never takers, and,

$$p(Y(1), R_y(1) \mid X, R_x, W, C = a) = p(Y(0), R_y(0) \mid X, R_x, W, C = a),$$

for always takers.

12.1.5 Noncompliance process assumption II – Monotonicity

Implicit in the definition of compliance status, C, and as pointed out in Section 11, we exclude the possibility that there exist people who will do the opposite of their assignment. These individuals are referred to in the compliance literature as "defiers" and have the property that, for individual i,

$$\begin{aligned} D_i(Z_i = 0) &= 1, \text{ and,} \\ D_i(Z_i = 1) &= 0. \end{aligned}$$

The assumption that there exist no defiers for this study is referred to as monotonicity because it implies that for all i, $D_i(Z_i = 1) > D_i(Z_i = 0)$ (Imbens and Angrist 1994). In the SCSFP study 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 highly implausible that such a group of people exists, therefore the monotonicity assumption appears to be quite reasonable.

12.2 Parametric Model

Consider the following factorization of the joint sampling distribution of the potential outcomes and compliance conditional on the covariates and their missing data patterns,

$$\begin{aligned} &p(Y_i(0), Y_i(1), R_{y_i}(0), R_{y_i}(1), C_i \mid W_i, X_i^{\text{obs}}, R_{x_i}, \theta) = \\ &p(C_i \mid W_i, X_i^{\text{obs}}, R_{x_i}, \theta^{(\text{C})}) p(R_{y_i}(0), R_{y_i}(1) \mid W_i, X_i^{\text{obs}}, R_{x_i}, C_i, \theta^{(\text{R})}) \\ &p(Y_i(0), Y_i(1) \mid W_i, X_i^{\text{obs}}, R_{x_i}, C_i, \theta^{(\text{Y})}) \end{aligned}$$

justified by the preceding assumptions. Note that the response pattern of covariates for each individual is itself a covariate.

The specifications of each of these components are described in the next three sections.

12.2.1 Compliance Status Sub-Model

The specification for the compliance status model comprises a series of 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, respectively:

$$C_i(n) = 1 \text{ if } C_i(n)^* \equiv g_1(W_i, X_i^{\text{obs}}, R_i)'\beta^{(\text{C},1)} + V_i \geq 0$$
$$C_i(c) = 1 \text{ if } C_i(n)^* < 0 \text{ and } C_i(c)^* \equiv g_1(W_i, X_i^{\text{obs}}, R_i)'\beta^{(\text{C},2)} + U_i \geq 0,$$

where

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

The specific models attempt to strike a balance between including all the design variables and the variables that were regarded as most important in predicting compliance or having interactions with the treatment effect, and on the other hand trying to maintain parsimony. The results reported in Section 13 use a compliance component model whose link function, g_1, fits, in addition to an intercept: the quality of school (bad/good); indicators for application wave; propensity scores for subjects applying in the first period; indicators for grade of the student; and an indicator for whether or not the pre-treatment test scores of reading and math were available; and the pre-test scores (reading and math) for the subjects with available scores.

Because the pre-tests were either jointly observed or jointly missing, one indicator for missingness of pre-test scores is sufficient. The same is true of the post-tests.

The prior distributions for the compliance sub-model are

$$\beta^{(\text{C},1)} \sim \text{N}(0, \{\sigma^{(\text{C},1)}\}^2\mathbf{I}),$$
$$\text{and} \quad \beta^{(\text{C},2)} \sim \text{N}(0, \{\sigma^{(\text{C},2)}\}^2\mathbf{I}),$$

where $(\sigma^{(\text{C},1)})^2$ and $(\sigma^{(\text{C},2)})^2$ are "known" hyperparameters set at five.

12.2.2 Outcome Sub-Model

The specification for the outcome sub-models is

$$Y_i(z) \mid W_i, X_i^{\text{obs}}, R_i, C_i, \theta^{(\text{Y})} \sim \text{N}(g_2(W_i, X_i^{\text{obs}}, R_i, C_i, z)'\beta^{(\text{Y})}, \exp[g_3(X_i^{\text{obs}}, R_i, C_i, z)'\zeta^{(\text{Y})}]),$$

for $z = 0, 1$, where $\theta^{(\text{Y})} = (\beta^{(\text{Y})}, \zeta^{(\text{Y})})$ and where $Y_i(0)$ and $Y_i(1)$ are assumed conditionally independent, an assumption which has no effect on inference for super-population parameters (Rubin 1978).

The results reported in Section 13 use an outcome component model whose outcome mean link function, g_2 is linear in, and fits distinct parameters for, the following:

1. For the students of the PMPD design: an intercept; the quality of school (bad/good); indicators for grade; the propensity score; and an indicator for whether or not the pre-treatment test scores were available, and the pre-test score values for the subjects with available scores.
2. For the students of the other periods: an intercept; the quality of school (bad/good); indicators for grade; indicators for application wave; an indicator for whether or not the pre-treatment test scores were available, and the pre-test score values for the subjects with available scores.
3. For students in the PMPD with observed pre-treatment score: an indicator for whether or not the person is a complier; and, when the person is complier, an assignment effect.
4. For students in all other waves with observed pretreatment score: an indicator for whether or not the person is a complier; and, when the person is complier, an assignment effect.
5. For students with missing pretreatment score: an indicator for whether or not the person is a complier; and, when the person is complier, an assignment effect.
6. An indicator for whether or not a person is an always-taker.
7. For compliers assigned treatment: one indicator for type of school (bad/good) and indicators for the the first three grades (the variable for the fourth grade's treatment effect is a function of the already included variables.)

For the variance of the outcome component, the link function, g_3, includes indicators that saturate the missing data patterns, which are defined by cross-classification of whether or not a person applied in the first wave (i.e., for whom there is a propensity score), and by whether or not the pre-treatment test scores were available. This dependence is needed because each pattern conditions on a different set of covariates; i.e., X^{obs} varies from pattern to pattern.

The prior distributions for the outcome sub-model are

$$\beta^{(\text{Y})} \mid \zeta^{(\text{Y})} \sim \text{N}(0, F(\zeta^{(\text{Y})})\xi\mathbf{I})$$
$$\text{and} \quad F(\zeta^{(\text{Y})}) = \frac{1}{K}\sum_k \exp(\zeta_k),$$

where $\zeta^{(\text{Y})} = (\zeta_1, \ldots, \zeta_K)$, one component for each of the K (in our case K=4) missing data patterns and where ξ is an "inflater" which is set at 5. In addition,

$$\exp(\zeta_k) \overset{\text{iid}}{\sim} \text{inv}\chi^2(\nu, \sigma^2),$$

where $\text{inv}\chi^2(\nu, \sigma^2)$ refers to the distribution of the inverse of a χ^2 random variable with degrees of freedom ν and scale parameter σ^2. Here ν is three and σ^2 is 300.

12.2.3 Outcome Response Sub-Model

We also use a probit specification for the sub-model for outcomes response, $R_{yi}(z)$, $z = 0, 1$.

$$R_{yi}(z) = 1 \text{ if } R_{yi}(z)^* \equiv g_2(W_i, X_i^{\text{obs}}, R_{xi}, C_i, z)'\beta^{(\text{R})} + E_i(z) \geq 0,$$

where $R_{yi}(0)$ and $R_{yi}(1)$ are assumed conditionally independent (using the same justification as for the potential outcomes) and where

$$E_i(z) \sim \text{N}(0, 1).$$

The link function of the probit model on the outcome response, g_3, is the same as the link function for the mean of the outcome component.

The prior distribution for the outcome response sub-model is

$$\beta^{(\text{R})} \sim \text{N}(0, \{\sigma^{(\text{R})}\}^2\mathbf{I}),$$

where $\{\sigma^{(\text{R})}\}^2$ is a "known" hyperparameter, set at five.

13 Results

In this section we first present answers to the three questions posed in Section 4

1. What is the impact of being offered a scholarship on student outcomes?
2. What is the impact of using a scholarship (participating in the scholarship program) over and above what families and children would do in the absence of the scholarship program?
3. What is the impact of attending a private school on student outcomes?

In all three cases math and reading post-test scores will be used as outcomes. These test scores represent the normal curve equivalents of 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; students transferring to private schools are hypothesized to be more likely to have been kept behind by those schools.

We also examine the distribution of compliance status and comparative missing of post-test scores between compliance groups across grade and school quality subgroups. In addition, we investigate the consistency in predicted compliance status across the two univariate (reading and math) analyses.

The following results are reported by school quality (bad/good) and grade, our "policy-relevant" variables, averaging over the other characteristics in the model. Both school quality and grade were thought to have possible interaction effect with treatment assignment. Most of the following estimates are not parameters of the model but functions of parameters, whose posterior distribution is induced by the posterior predictive distribution (multiple imputation) of the compliance status. Except when otherwise stated, plain numbers are posterior means and brackets are 95% posterior intervals.

13.1 ITT results

We examine the impact of being offered a scholarship on post-test scores by estimating the ITT effect as displayed in Table 11.

These results indicate posterior distributions primarily (95%) to the right of zero for the treatment effect on reading scores for 4th graders from bad schools and on math

Grade at entry	Reading		Mathematics	
	Good School	Bad School	Good School	Bad School
1	0.43 [-2.30, 3.10]	1.20 [-0.68, 3.22]	3.10 [0.84, 5.46]	4.49 [2.67, 6.45]
2	-2.45 [-5.07, 0.07]	-1.21 [-2.99, 0.58]	-0.86 [-3.66, 1.35]	0.55 [-1.29, 2.24]
3	-0.67 [-3.02, 1.83]	0.32 [-1.20, 1.77]	1.75 [-0.63, 4.08]	3.38 [1.53, 5.00]
4	0.26 [-3.29, 3.56]	3.30 [1.15, 5.41]	-0.06 [-2.60, 2.77]	3.14 [1.35, 5.19]

Table 11: ITT Effect

scores for 1st, 3rd, and 4th graders from bad schools. All but one of the intervals for children applying from good schools cover zero.

13.2 Analysis with never takers

The first analysis will allow for never takers but not always takers. That is, we will measure the ITT effect for always takers and compliers combined. This analysis defines the SCSFP as the "treatment" rather than just private school attendance. This will provide an answer to the second of the questions posed above because the complier control group will include children who were able to take advantage of resources beyond those provided by the SCSFP.

We seem a similar pattern of effects in this analysis (with the addition of a "significant" effect on math test scores for 1st graders from good schools) though the posterior means corresponding to intervals which don't cover zero are all larger than in the ITT analysis. The intervals are all larger than the ITT intervals which is not surprising given that the estimand now applies to only a subset of the study participants.

Grade at entry	Reading		Mathematics	
	Good School	Bad School	Good School	Bad School
1	0.61 [-3.50, 4.41]	1.63 [-0.87, 4.34]	4.56 [1.21, 8.06]	6.19 [3.59, 9.78]
2	-3.49 [-7.32, 0.11]	-1.52 [-3.82, 0.69]	-1.23 [-5.11, 2.30]	0.71 [-1.57, 2.91]
3	-0.97 [-4.43, 2.68]	0.43 [-1.65, 2.40]	2.69 [-0.99, 6.40]	4.65 [2.16, 7.09]
4	0.38 [-5.03, 5.31]	4.22 [1.53, 6.68]	-0.05 [-4.00, 4.28]	4.06 [1.72, 7.03]

Table 12: Effect of Scholarship Program

13.3 Analysis with never takers and always takers

The second analysis will allow for never takers as well as always takers. Children who were assigned to the control group but attended private school will be coded as $D_i = 1$. This analysis defines the "treatment" as private school attendance. The validity of the estimand "effect of private school attendance" depends on the assumption that receiving a scholarship and then attending private school is the same treatment as not receiving a scholarship and attending private school. In addition, it ignores the fact that children who received a scholarship were provided with help from the SCSF in finding an appropriate private school; those who didn't receive a scholarship but nevertheless attended private school were provided with no such assistance.

The effects of private school attendance, displayed in Table 13 are quite similar to the scholarship program effects which again slightly higher means for those subgroups whose posterior intervals do not cover zero.

13.4 Composition of compliance status

Table 14 gives estimates of the composition of compliance status as a function of school quality and grade. Because the distributions between the two models (mathemat-

Grade at entry	Reading		Mathematics	
	Good School	Bad School	Good School	Bad School
1	0.63 [-4.04, 4.57]	1.83 [-0.97, 4.89]	4.97 [1.27, 8.91]	6.91 [4.10, 9.89]
2	-3.89 [-9.12, 0.12]	-1.73 [-4.43, 0.77]	-1.41 [-5.84, 2.65]	0.84 [-1.84, 3.50]
3	-1.09 [-5.17, 2.92]	0.48 [-1.90, 2.68]	3.08 [-1.06, 7.36]	5.32 [2.48, 8.08]
4	0.39 [-5.79, 6.06]	4.70 [1.78, 7.52]	-0.04 [-4.32, 4.90]	4.55 [1.99, 7.88]

Table 13: Effect of Private School Attendance

ics/reading) were comparable in both location and uncertainty (see also Section 13.6 below), reported results are from the equal-weight mixture of the distributions of the two models.

The clearest pattern revealed by Table 14 is that good schools have more never takers and fewer compliers than bad schools.

13.5 Impact of missing data

When the latent compliance groups have differential response behaviors, standard ITT analyses or standard IV analyses are generally not appropriate for estimating, respectively, the ITT or IV estimands. The following table compares response behavior (i) between compliers attending public schools and never-takers, (ii) between compliers attending private schools and always takers, and (iii) between compliers attending private schools and compliers attending public schools.

The observed response behavior on the mathematics and reading was identical within individuals. For this reason, and also because, there was satisfactory agreement in the prediction of compliance status between the two models (mathematics/reading, see Section 13.6 below), reported results are from the equal-weight mixture of the distributions from the two models. In addition, the posterior distributions of the odds ratios

Grade	School	Never Taker	Complier	Always Taker
1	Good	30.9 (6.0)	63.7 (6.7)	5.4 (3.0)
	Bad	26.7 (3.6)	65.6 (4.3)	7.7 (2.4)
2	Good	31.1 (6.3)	60.3 (8.2)	8.6 (4.5)
	Bad	21.4 (3.5)	68.0 (5.2)	10.6 (3.3)
3	Good	32.4 (6.3)	60.1 (7.5)	7.5 (3.6)
	Bad	26.3 (3.8)	64.7 (4.5)	9.0 (2.4)
4	Good	34.4 (6.8)	59.5 (7.6)	6.1 (3.1)
	Bad	22.1 (4.1)	69.7 (5.1)	8.2 (2.9)

Posterior standard deviations are in parentheses.

Table 14: Composition of compliance status

are skewed, so posterior medians and posterior intervals are reported.

For each of the first two comparisons, the groups being compared are attending the same type of school, so any difference in response rate is attributed to the latent compliance status characteristics. For the last comparison, any differences are attributed to the treatment. From the table it can be deduced that response is higher in the following order: never-takers, compliers attending public, compliers attending private, and always-takers. Therefore, the latent compliance behavior is an important predictor of response.

13.6 Agreement between the models

We assess the agreement between the two models in predicting compliance type (i) at the individual student level, and (ii) as a function of the covariates bad/good and grade, aggregating over the students in these classes. Evaluating agreement at such specific levels is important because, although the marginal probability of being a complier is well estimated generally, the two models might have been assigning different probabilities of being a complier to different sets of students.

For the individual level, for each model, and for each student assigned the lottery but

Grade	School	Never Taker	Complier	Always Taker
1	Good	2.0 [1.0,3.9]	0.5 [0.0,1194.5]	2.6 [0.7,1002.4]
	Bad	2.1 [1.1,3.9]	0.3 [0.0, 2.3]	1.4 [0.6, 3.1]
2	Good	1.9 [0.9,3.9]	0.5 [0.0, 662.9]	3.9 [1.1, 833.8]
	Bad	2.1 [1.1,3.9]	0.4 [0.0, 189.8]	2.9 [1.1, 245.4]
3	Good	2.2 [1.1,4.3]	0.3 [0.0, 903.2]	2.8 [0.6,1177.9]
	Bad	2.2 [1.2,4.1]	0.3 [0.0, 4.3]	1.6 [0.7, 4.4]
4	Good	2.6 [0.9,5.8]	0.3 [0.0, 982.7]	2.3 [0.5, 951.4]
	Bad	2.1 [1.1,3.9]	0.3 [0.0, 3.4]	1.6 [0.7, 4.3]

Odds Ratios of $R_{ij} = 1$ vs. $R_{ij} = 0$

Table 15: Composition of compliance status

whose compliance type was not known, we compute the posterior probability of being a complier. The correlation between the probabilities obtained from the two models was 0.72, and the corresponding correlation for the students assigned control with unknown compliance status was 0.73, indicating a satisfactory level of agreement at the individual level. At the level of the cross-classification between grade and bad/good the agreement of the posterior distributions of compliance status, summarized by posterior first two moments, was very good.

14 Discussion

Results from these analyses, which rely on far weaker assumptions than those from the complete case analyses, lead to different inferences than those indicated by the complete case analyses. In all cases, math seems to be the area where we see improvement, not reading, and there is a clear interaction between the treatment in each analysis and the quality of the school the children attended prior to the study.

Acknowledgments

David Myers and Paul E. Peterson were co-principal investigators for the evaluation. We wish to thank the School Choice Scholarships Foundation (SCSF) for co-operating fully with this evaluation. This evaluation has been supported 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. We are grateful to Kristin Kearns Jordan and other members of the SCSF staff for their co-operation and assistance with data collection. We received helpful advice from Paul Hill, Christopher Jencks, and Donald Rock. Daniel Mayer and Julia Kim, from Mathematica Policy Research, were instrumental in preparing the survey and test score data and answering question about that data. Additional research assistance was provided by David Campbell and Rachel Deyette; staff assistance was provided by Shelley Weiner. 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.

This paper is, in part, an amalgam of other articles. The introduction and portions of Sections 3 and 5 were taken from Peterson and Howell (1999). The Design Section is a slightly modified version of portions of Hill, Thomas, and Rubin (1999).

References

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

Ascher, C., Fruchter, N., and Berne, R. (1996), "Hard Lessons: Public Schools and Privatization," Tech. rep., Century Foundation, New York, NY.

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.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analyses: Theory and Practice*, MIT Press.

Bonsteel, A. and Bonilla, C. A. (1997), *A Choice for our Children: Curing the Crisis in America's Schools*, San Francisco, California: Institute for Contemporary Studies.

Brandl, J. E. (1998), *Money and Good Intentions are not Enough, or Why Liberal Democrat Thinks States Need Both Competition and Community*, Washington, D.C.: Brookings Institution Press.

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

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

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

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya* 35, 417–446.

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

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

Coulson, A. J. (forthcoming), "Market Education: The Unknown History," .

D'Agostino, Ralph B., J. and Rubin, D. B. (1999), "Estimating and Using Propensity Scores With Incomplete Data," pending publication in JASA.

Derek, N. (1997), "The Effects of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics* 15, 1, 98–123.

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.

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.

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, 103–22.

Gu, X. S. and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics* 2, 405–420.

Gutmann, A. (1987), *Democratic Education*, Princeton, NJ: Princeton University Press.

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

Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Econometrica* 12, 1–115, (Supplement).

Hill, J. L., Thomas, N., and Rubin, D. B. (1999), "The Design of the New York School Choice Scholarship Program Evaluation," in *Donald Campbell's Legacy*, ed. L. Bickman, Sage Publications.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, A. (1999), "Estimating the Effect of an Influenza Vaccine in an Encouragement Design," unpublished.

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. (1997a), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *The Annals of Statistics* 25, 305–327.

Imbens, G. W. and Rubin, D. B. (1997b), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies* 64, 555–574.

Levin, H. M. (1998), "Educational Vouchers: Effectiveness, Choice, and Costs," *Journal of Policy Analysis and Management* 17, 3, 373–392.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: John Wiley & Sons.

Meng, X.-L. and Rubin, D. B. (1993), "Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework," *Biometrika* 80, 267–278.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Chemical Physics* 21, 1087–1091.

Mosteller, F. (1995), "The Tennessee Study of Class Size in the Early School Grades," in *The Future of Children*, vol. 5, pp. 113–27.

Muthen, L. K. and Muthen, B. O. (1998), *Mplus: The Comprehensive Modeling Program for Applied Researchers. User's Guide*, Muthen & Muthen.

Neyman, J. (1923), "On the Application of Probablity Theory to Agricultural Experiments Essay on Principles. Section 9," translated in *Statistical Science* 5, 465–480, 1990.

Over, M., Jolliffe, D., and Foster, A. (1995), "Huber correction for two-stage least squares estimates," Tech. rep., Stata Technical Bulletin, Reprinted in Stata Technical Bulletin Reprints, vol.5, p.140-142.

Peterson, P. E. and Hassel, B. C., eds. (1998), *Learning from School Choice*, Washington, D.C.: Brookings Institution Press.

Peterson, P. E. and Howell, W. G. (1999), "What Happens to Low-Income New York Students When They Move from Public to Private Schools," in *City Schools: Lessons from New York*, eds. D. Ravitch and J. Viteritti, Johns Hopkins University Press, forthcoming.

Rasell, E. and Rothstein, R., eds. (1993), *School Choice: Examining the Evidence*, Washington, D.C.: Economic Policy Institute.

Roseman, L. (1998), "Reducing Bias in the Estimate of the Difference in Survival in Observational Studies Using Subclassification on the Propensity Score," Ph.D. thesis, Harvard University.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70, 1, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79, 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician* 39, 33–38.

Rubin, D. B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics* 29, 185–203.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology* 66, 688–701.

Rubin, D. B. (1977), "Assignment to Treatment Groups on the Basis of a Covariate," *Journal of Educational Statistics* 2, 1–26.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The role of randomization," *The Annals of Statistics* 6, 34–58.

Rubin, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74, 318–328.

Rubin, D. B. (1980), "Comments on "Randomization Analysis of Experimental Da ta: The Fisher Randomization Test"," *Journal of the American Statistical Association* 75, 591–593.

Rubin, D. B. (1990), "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science* 5, 472–480.

Rubin, D. B. and Thomas, N. (1992), "Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions," *Biometrika* 79, 797–809.

Rubin, D. B. and Thomas, N. (1996), "Matching using estimated propensity scores: Relating theory to practice," *Biometrics* 52, 249–264.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

StataCorp. (1997), *Stata Statistical Software: Release 5.0*, College Station, TX: Stata Corporation.

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.

Appendix A – Computations

Computations of the posterior distribution of the missing compliance statuses C^{mis} and parameters were based on a Gibbs sampler (Gelfand and Smith 1990). The Gibbs sampler we used draws, in this order: the missing compliance statuses C^{mis}; the latent variables $C_i(n)^*$ and $C_i(c)^*$ for the current set of never takers, compliers, and always-takers; the latent variables $R_{yi}(Z_i)^*$ for the response model; the parameters for the compliance model, $\beta^{(c,1)}$, $\beta^{(c,2)}$; the response model parameters $\beta^{(R)}$; and the mean and variance outcome parameters $\beta^{(Y)}$ and $\zeta^{(Y)}$ respectively. For all steps drawing is done cyclically and each step except the first conditions on all other unknowns by setting their values to those obtained from the most recent cycle. The first step must exclude $C_i(n)^*$ and $C_i(c)^*$ from the conditioning in order for the Gibbs sampler to converge to the posterior. In the following we let $H_i \equiv (W_i, X_i^{obs}, R_{xi})$ and ϕ denote all of the model parameters. The distributions involved in the Gibbs sampler are as follows.

1. The conditional distribution required for C_i^{mis} at this step is

$$p(C_i|Y_i, H_i, D_i, Z_i, R_{yi}, \phi).$$

   This distribution is obtained from the joint $p(C_i, Y_i, H_i, D_i, R_{yi}|Z_i, \phi)$. For example, a subject with $Z_i = D_i = 0$ can be a complier or a never-taker, and the conditional Bernoulli distribution of C_i is proportional to

$$\{l(c, Z_i, H_i, Y_i, R_{yi}, \phi)\}^{I(C_i=c)}\{l(n, Z_i, H_i, Y_i, R_{yi}, \phi)\}^{I(C_i=n)},$$

   where we define

$$\begin{aligned} l(c_0, z_0, h_0, y_0, r_0, \phi) = \quad & p(C_i = c_0|H_i = h_0, \phi)\{p(Y_i = y_0|C_i = c_0, H_i = h_0, z = z_0, \phi)\}^{r_0} \\ & \times p(R_{yi}(Z_i) = r_0|C_i = c_0, H_i = h_0, \phi). \end{aligned}$$

   Therefore, the conditional probability of the subject being a complier is

$$l(c, Z_i, H_i, Y_i, R_{yi}, \phi)\{l(c, Z_i, H_i, Y_i, R_{yi}, \phi) + l(n, Z_i, H_i, Y_i, R_{yi}, \phi)\}^{-1}.$$

   The drawing of C_i for subjects with $Z_i = D_i = 1$ is done in a similar way. Note that the drawing of the compliance at this step uses information on the response behavior (R_y).

2. The drawing of $C_i(n)^*$ is from $p(C_i(n)^*|H_i, C_i, \phi)$. This distribution is the same as the defining model $p(C_i(n)^*|H_i, \phi)$ but truncated either to the left or to the right of zero depending on C_i. The drawing of the truncated normal is done using its inverse distribution function, which is readily calculable. For subjects that have been imputed as always-takers or compliers at the previous step, drawing of $C_i(c)^*$ is done in a similar way.

3. The drawing of $R_{yi}(Z_i)^*$ is from $p(R_{yi}(Z_i)^*|H_i, C_i, R_{yi}, z = Z_i, \phi)$. This distribution is the same as the defining model $p(R_{yi}(Z_i)^*|H_i, C_i, z = Z_i, \phi)$ except that it is truncated to the right or left of zero depending on R_{yi}. Drawing is as with the compliance latent normals.

4. Drawing of the coefficients $\beta^{(c,1)}$ is from $p(\beta^{(c,1)}|\{$ all $C_i(n)^*, H_i\})$, which is a Bayesian linear regression based on the defining likelihood and prior. Drawing of the coefficients $\beta^{(c,2)}$ is from $p(\beta^{(c,2)}|\{$ all $C_i(c)^*, H_i : C_i = a$ or $c\})$, and drawing of the coefficient $\beta^{(R)}$ is from $p(\beta^{(R)}|\{$ all $R_{yi}(Z_i)^*, H_i, Z_i, C_i\})$, both of which are Bayesian linear regressions.

5. The drawing of the parameters of the outcome model is further divided in two steps. In one, with $\zeta^{(Y)}$ conditioned at the values from the previous cycle, $\beta^{(Y)}$ is drawn from $p(\beta^{(Y)}|\{$ all $Y_i, H_i, Z_i, C_i : R_{yi} = 1\}, \zeta^{(Y)})$, which is a weighted normal linear regression with known weights. With the mean parameters $\beta^{(Y)}$ conditioned at the drawn value, there is still no known direct way of drawing from the distribution of $\zeta^{(Y)}$. Nevertheless, because its distribution is easily calculable up to proportionality, the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) was used. Because the dimension of the parameters is large, it is important to obtain a good jumping density. By defining, $Y_i^* = |Y_i - g_2(H_i, C_i, Z_i)\beta^{(Y)}|$ for the observed outcomes we have that

$$\frac{(Y_i^*)^2}{\exp(g_3(H_i, C_i, Z_i)\zeta^{(Y)})} \sim \chi_1^2, \quad \text{so}$$
$$E(2\log(Y_i^*)) = E(\log(\chi_1^2)) + g_2(H_i, C_i, Z_i)\zeta^{(Y)},$$

where χ_1^2 is a chi-squared random variable with one degree of freedom and the distribution and expectation above are conditional on all variables except Y_i and

parameters including $\zeta^{(Y)}$. Using the last regression we obtain two moments that we used in a normal jumping density for $\zeta^{(Y)}$. Because the jumping density does not use the values of $\zeta^{(Y)}$ from the previous cycle, the asymmetric version of Metropolis-Hastings was used.

Initial values for the missing compliance statuses were drawn based on the moment estimates given assignment arm and school attended. The parameters were initialized to generalized linear model estimates given the initialized compliance statuses. Subsequently, the models were run each for an initial burnout series of 1500 iterations. We assessed convergence with add hoc methods. Then a main series of an additional 5500 iterations was run for each model, on which the results are based.

Appendix B

Table 16: Description of Variables

Variable	Description
Baseline variables (pre-lottery)	
Application wave	Indicator for each of five waves
Won a scholarship?	No/Yes
Bad/good school	Indicator for each category
Child's birth location	U.S./Puerto Rico/Other
Grade level of child when applying	Kindergarten through 4th grade
Female guardian's ethnicity	Puerto Rican, Dominican, Other Hispanic/Black,African American/Other
Female guardian's education	Some high school/High school graduate or GED/Some college/Graduated from a 4 year college/More than a 4 year degree
Child participated in special education in the last year?	No/yes
Child participated in gifted programs in the last year	No/yes
Main language spoken in home	English/Other
Family participates in AFDC	Yes/No
Family participates in Foodstamp Program	Yes/No
Female guardian's work status	Fulltime/Part-time/Not working but looking/Not working not looking
Education expectations for child	Some high school will not graduate/Graduate from high school/Some college/Graduate from 4-year college/More than a 4-year college degree
Number of children under 18 in household	
Female guardian's birth location	United States/Other
Female guardian's length of residence at current address	More than 2 years/1-2 years/3-11 months/Less than 3 months
Data on father's work status missing?	No/Yes
Female guardian's religion	Other/Catholic
Sex	Male/Female
Income	0-$4999/$5000-7999/.../More than $50,000
Age of the child on 4/1/97 in years	
Pre-test reading score (percentile)	
Pre-test math score (percentile)	
continued on next page	

continued from previous page	
Variable	Description
Pre-test reading score (normal curve equivalent)	
Pre-test math score (normal curve equivalent)	
Attendance at private school during previous year	No/Yes
Survey respondent one of child's primary caretakers what portion of the time during the past year?	None/Some/All
Time student has attended day care/school outside the US?	None/Some
Where send child to school next year (if no scholarship)?	Public/Religious Private/Secular Private
How many times during the school year have you spoken to someone from this child's school about "problems with this child's' behavior at school"?	None/1 or 2/3 or 4/More Than 4
How many times during the school year have you spoken to someone from this child's school about "this child's attendance"	None/1 or 2/3 or 4/More than 4
How many times have during the school year have you spoken to someone from this child's school about "placing this child in special classes or programs"	None/1 or 2/3 or 4/More than 4
Variables recorded one year after the lottery	
Post-test reading score (percentile)	
Post-test math score (percentile)	
Post-test reading score (normal curve equivalent)	
Post-test math score (normal curve equivalent)	