The paper has been accepted and will apear in 2001 in: *Case Studies in Bayesian Statistics,* C. Gatsonis et al. (eds.). New York: Springer-Verlag, with discussion by Stephen Ansolabehere (MIT), Brian Junker (Carnegie Mellon University), and Alix Gitelman (Oregon State University).

School Choice in NY City: A Bayesian Analysis of an Imperfect Randomized Experiment

John Barnard Constantine Frangakis Jennifer Hill Donald B. Rubin

February 23, 2001

Abstract

The precarious state of the educational system existing in the inner-cities of the U.S., 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 analysis of data from 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. However, this program evaluation, as is common in studies with human subjects, suffers from unintended, although not unanticipated, 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 parents failed to complete fully survey information; some children did not 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 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.

These analyses provide evidence of positive effects of private school attendance on math test scores for certain subgroups of the children studied.

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 can be 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, this should not lessen the statistician's responsibility to model, as rigorously as possible, the science of the problem.

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 defined broadly 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 frame these complications as aspects of the broadly defined phenomenon of interest and then build them into our Bayesianly-inspired template for the study and its data.

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

In this way, this application may stand in contrast to many Bayesian applications that 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 influenced

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 them into the analysis (for instance inclusion of survey questions intended to help the modeling of these complications).

Our analysis does not represent a completely satisfactory job of simultaneously handling all the complications. As an example, we do not model the multivariate nature of the outcomes; we fit separate models for each outcome examined (reading and math test scores). Further work will gradually expand this initial model to incorporate the complicated structure of this experiment and the "response" to school choice that it measures. The more the model becomes more inclusive of the complications, the more we will be able to take advantage of the elements we incorporated in our initial design that anticipated this structure.

2 Introduction

Over the past few years, interest in school choice 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 arguments both for and against the continuation and/or expansion of these school choice programs. Supporters of school choice assert that low-income, inner-city children learn more in private schools; critics retort that any perceived learning gains in pri-

vate schools are due to the selected nature of private-school families. Proponents suggest 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 very few voucher experiments have been attempted. 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 insurmountable. The best solution is to implement numerous large-scale controlled randomized experiments. 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 on test scores 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-

[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).

funded school choice program. The SCSF 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 collected 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 program 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 imperfect randomized experiment and the corresponding notation is given in Section 9. The model is described in Section 10; technical details of the computations are reserved for Appendix A. Results of the analysis are discussed in Section 11.

3 School Choice Scholarships Foundation (SCSF) Program

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 program 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 city-wide median (henceforth labeled "low-score" schools). 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 conducted by MPR; the co-principal investigators were David Myers, MPR, and Paul Peterson, Harvard University (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? Each of these questions may be answered by using information collected for the SCSF evaluation. Until this evaluation, no one study has addressed these three questions. Furthermore, this study may 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 children's education, and student test scores, both prior to the lottery and one year later; 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 as will be discussed in greater detail in the following sections.

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 the perfectly controlled experiment effected by families and children. We have identified four such behaviors:

1. Some families offered a scholarship did not subsequently accept the scholarship and attend a private school.
2. Some families not offered a scholarship sent their children to a private school anyway[3].
3. Some families invited to attend data collection and testing sessions one year after the baseline survey did not show up.
4. Some parents and students did not complete all items in their questionnaire, and some 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 "non-compliance," the last two as "missing data." For the SCSF evaluation, ensuring compliance

[3]Classifying this behavior as a deviation assumes that the treatment is defined as private school attendance and that the range of private schools attended by the treatment group is similar to the private schools attended by the control group. This issue will be discussed in greater detail in Section 11.

with the assigned treatment was largely out of the control of the evaluation team. If we define treatment as private school attendance, clearly the team could neither force winners to use their scholarships, nor could they keep those who did not win from attending private school. The SCSF did, however, provide services to help scholarship winners find appropriate private schools, which may have helped compliance rates. If we define treatment as participation in the scholarship program, then the only form of non-compliance is scholarship winners deciding not to participate in the program (clearly those who did not win could not obtain a scholarship or receive help from program administrators in finding a school). Again, provision of help in finding private schools for scholarship winners probably may have lessened non-compliance.

In social science studies it is generally difficult for evaluators to have much control over noncompliance of the control group with respect to participating in program services. They cannot prevent 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. It is also unclear that evaluators should want to prevent such actions. If we want the study to answer a public policy question (e.g. "Should we make available Program A? Will it make a difference in this community?"), the correct control should probably represent the other services the target population has available to them. However, in this case the issue is often "Do students learn more in private schools?".

Evaluators generally have more control, potentially, over the amount or kinds of missing data that 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 did not take the test (see Section 5.5). 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 between the evaluation team and some of the authors 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. This was done in anticipation of structuring non-compliance and missing data into our eventual Bayesian analysis.

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 and staff 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 private 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. To preserve the integrity of the control group, we[4] randomly selected about 100 winners for the second year of scholarships. Second, a variable incentive scheme allowed many control group families that attended a testing session to receive an incentive of $75 on average (some were offered $50 and others were offered $100).

[4]This was actually performed by colleague Neal Thomas, see Hill, Rubin, and Thomas (1999).

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 were obtained for applicants in kindergarten because: (1) these children would most likely never have been exposed to a standardized test hence considerable time would have been spent instructing the children on how to take a test, and (2) there was concern that separating such young children from their guardians in this new environment with unfamiliar teachers might lead to discipline or behavioral issues. 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

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

differently. Second, we do not yet have complete compliance data for the multi-child families. For this reason, the analyses in this paper are limited to results for the 1250 "single-child" families (i.e. families that only had one child participating in the lottery) who were in grades 1-4 at the time of the spring of 1997 application process.

6 Design

Although the lottery used to award scholarships naturally created a randomized design, it also precluded blocking on variables selected purely for their assumed predictive power. Randomization within certain subgroup classifications (e.g. ethnicity) might have appeared inequitable to the public[5]. Another complication was that evaluation funding only allowed for 1000 treatment families and 1000 control families to be followed. How to choose the control families from the reservoir of over 4000 families who participated in the lottery but did not 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

[5]The randomization was slightly constrained as will be described in more detail in this section. However, in one case this was done to ensure higher representation from a more disadvantaged population, and this policy was clearly stated in advertisements for the program. The other "blocks" –application wave and family size – were present for logistical reasons concerning data collection and allocation of a fixed number of scholarships. In general, in this type of program, administrators would like to keep these types of deviations from a pure lottery to a minimum.

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 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 program, applicants from "low-score" schools (schools whose average test scores were below the city-wide median) were given a higher chance of winning a scholarship than those from "high-score" schools (schools whose average test scores were above the city-wide median). Families from "low-score" 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 high versus low-score 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 low/high 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 low/high variable as well as variables correlated with this variable.

6.2 Multi-child Families

The SCSF program 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., low/high 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 low/high 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 children) and

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

design (PMPD vs. randomized block). The initial sample sizes, [6] 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.

[6] 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 low versus high-test-score 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 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

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

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.

Methods for estimating propensity scores 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, for which, 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[7]. Mahalanobis matching is used to choose a "nearest neighbor" within this subset of study 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.

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

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 (in Section 6.5). 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, work to 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 SCSF program 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 SCSF program 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.[8].

The Mahalanobis matching within propensity score calipers in the SCSF program 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 low/high 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.

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), a simple random sample chosen from the control reservoir in the first time period, and a stratified random sample also chosen from the control reservoir in the first time period. The stratified random sample was randomized within low/high school categories; 85% of the children were chosen to be from low-score schools and 15% from high-score 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.

[8]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).

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), low/high 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 low/high 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 low/high), 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 predictive of outcomes[9] 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[10]. 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 low/high school) 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 5. It is beaten by the simple random sample for 6 variables and by the stratified random sample for 9 variables (there is one tie). In addition, the gains when PMPD beats its competitors are generally larger than the gains of the competitors over PMPD. The superior performance of the stratified random sample also reflects the gains which can be made when a control reservoir of this size is available for choosing the control group to be followed.

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 standard-

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

[10]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.

	Application Wave 1			Waves 2-5
Variable	Simple Random Sample	Stratified Random Sample	PMPD	Randomized Block
low/high	-0.98	0.00	0.11	0.21
grade level	-1.63	0.03	-0.03	-0.39
reading score	-0.38	0.65	0.48	-1.05
math score	-0.51	1.17	0.20	-1.37
ethnicity	1.80	1.68	1.59	1.74
mom's education	0.16	0.14	0.09	1.67
special education	0.31	1.66	-0.17	0.22
gifted program	0.42	-1.16	-0.13	0.75
language	-1.06	-0.02	-1.03	-0.44
afdc	-0.28	0.49	0.83	-1.57
food stamps	-1.08	-0.27	0.94	-1.31
mother works	-1.26	-0.30	-1.18	0.40
educ. expectations	0.50	1.79	0.57	0.19
children in household	-1.01	-1.75	0.41	-1.02
birth location	0.49	0.73	-1.40	-0.69
length of residence	0.42	0.71	0.66	-0.78
dad's work missing	1.09	0.70	0.00	0.16
religion	-1.84	-0.19	-0.74	-0.80
sex	0.88	1.22	0.76	0.53
income	-0.38	-0.62	0.74	-1.21
age as of 4/97	-1.57	0.18	-0.47	-0.87

Table 6: Balance: Single-Child Families

ized 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.

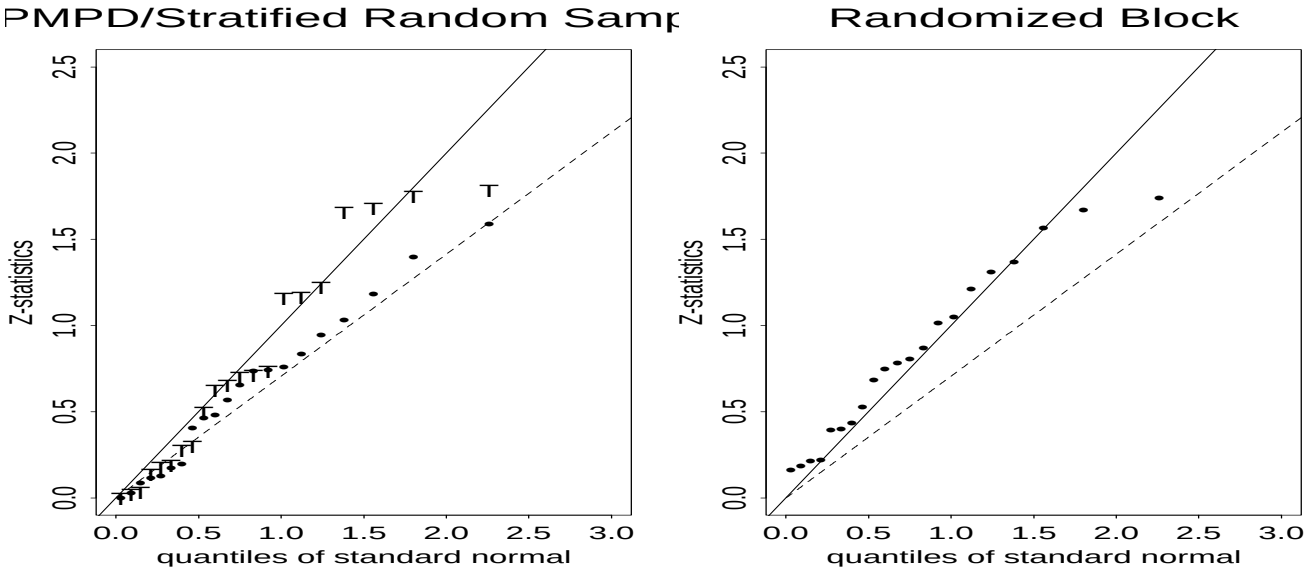

Figure 1: Half-Normal Plots of Z-Statistics for Single-Child Families

For Figure 1, the solid line in each panel corresponds to the normal distribution with variance 1 and the the dotted line in each panel corresponds to the normal distribution with variance $1/2$. The dots in the left and right panels represent the Z-statistics from the PMPD and randomized block designs respectively. This figure thus reveals that the gains predicted by Rubin and Thomas (1992) for the propensity score matching are fairly closely achieved for the study of single-child families. These results can be contrasted with those from the randomized block experiment, which are consistent with the standard normal distribution. The stratified

random sample (displayed as "T" points) is the best of the alternatives but still fails to achieve the efficiency gains of the PMPD. We excluded the simple random sample as it is an unlikely alternative given the initial low/high imbalance in the first application wave.

Since the variance in the difference in means is reduced by a factor of two, this is equivalent, for some analyses, to increasing the sample size by a factor of two for these variables. This principle holds, for instance, for any linear combination of the measured covariates, however, 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, low/high, 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 low/high 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 low/high variable is also present with the multi-child families, but the PMPD still achieves very good overall balance. Compared to the all other designs, the PMPD has lower absolute z-scores for 18 variables, higher z-scores for 8. Again, the gains when PMPD beats the other designs are generally larger than the other way around.

Half-normal quantile-quantile plots for the multi-child families in both experiments, dis-

	Application Wave 1			Waves 2-5
Variable	Simple Random Sample	Stratified Random Sample	PMPD	Randomized Block
low/high	-3.81	0.00	-0.98	0.15
avg. grade level	-0.27	0.21	0.38	0.23
s.d. grade level	-0.19	-0.08	-0.40	0.58
avg. reading score	-1.06	0.97	0.91	-0.23
s.d. reading score	-0.90	-1.95	1.23	-2.20
avg. math score	-0.56	0.26	0.82	0.32
s.d. math score	-1.02	-1.23	0.33	-1.11
ethnicity	-1.03	-0.95	0.20	2.09
mom's education	-0.27	-1.01	-0.21	-0.22
special education	-0.67	-1.12	-0.11	0.68
gifted program	-0.85	0.43	-0.07	-0.52
language	1.13	1.35	0.92	-0.64
afdc	-1.24	0.00	0.13	3.42
avg. age	-0.38	-0.19	0.48	0.66
s.d. age	0.09	0.14	0.00	0.38
avg. educ. exp.	-0.81	-1.22	0.49	-0.71
s.d. educ. exp.	-1.59	-0.80	-0.10	0.94
children in household	0.39	-0.27	-0.40	-0.13
income	0.93	1.47	0.13	2.01
religion	0.01	0.93	-0.97	-0.66
length of residence	-1.29	-1.44	0.54	1.31
dad's work missing	0.39	-1.91	0.70	1.73
food stamps	-2.06	-0.42	-0.35	2.58
mom works	1.29	0.87	0.73	-0.49
birth	0.20	1.26	-0.42	1.34
sex	-0.84	-0.07	-0.17	-1.43

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

played 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. The stratified random sample performs slightly better than the other alternatives but fails once again to achieve the efficiency of the PMPD.

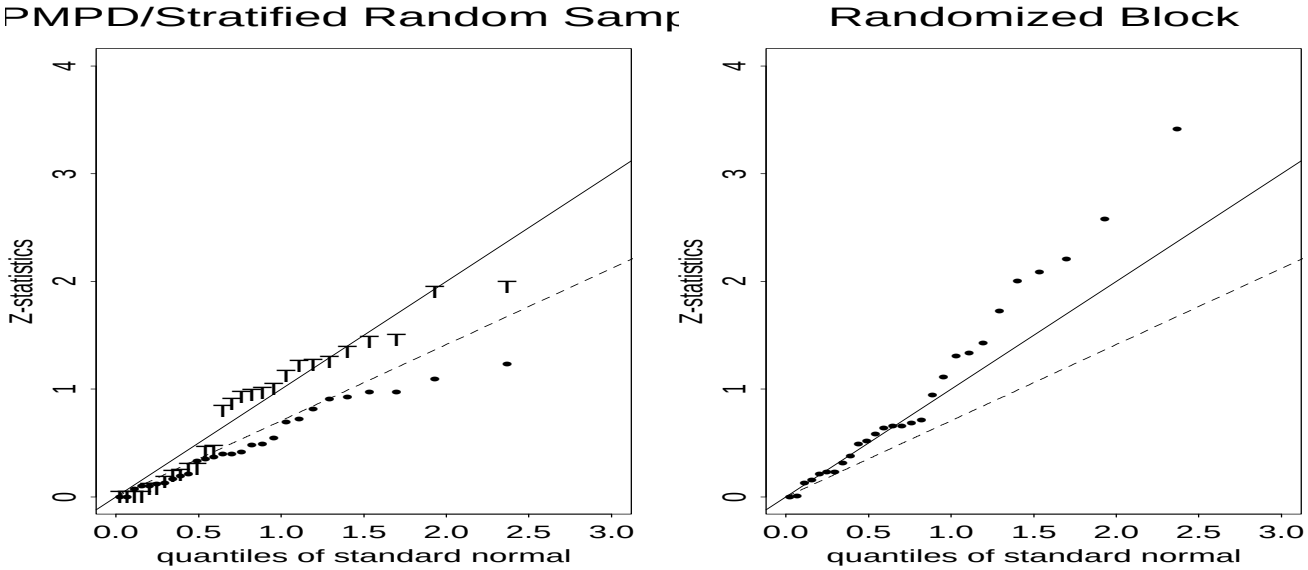

Figure 2: Half-Normal Plots of Z-Statistics for Multi-Child Families

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 low/high 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.

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 Imperfect Randomized Experiments

It is important to realize that our randomized experiment does not really randomize the treatment of, for instance, public and private school attendance 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 typically would be 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 1978a, 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 (1997) showed how the Bayesian approach to causal inference in Rubin (1978a) could be extended to handle simple randomized experiments with noncompliance, 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, and defining the treatment as private school attendance, will attend private school whether or not they are 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 (1997) 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 (1997) 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 reasonable interval, e.g. a 90% interval, and not whether or not an $\epsilon\%$ interval is unique as positive $\epsilon \to 0$.

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 a method of moments 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 Original MPR Analysis Strategy

Before introducing our Bayesian model we first present results from an analysis that combines several existing approaches to each of the three major complications we have discussed. This analysis strategy can be implemented with available software. Missing covariates are handled by limiting the number of covariates to the design variables and the most important predictors, the pre-test scores, and then including in the analysis only individuals for whom these variables are fully observed. Missing outcomes are adjusted for by non-response weights. Instrumental variable models are used to handle the non-compliance. Separate analyses are run for math scores and reading scores (national percentile rankings). Weights are used to make the results for the study participants representative of the population of all eligible single-child families who were screened. The results in this section are obtained using the same analysis strategy that was used in the initial MPR study (Peterson, Myers, Howell, and Mayer 1999) but now only on the subset of single-child families and separated out by the low/high variable.

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	-0.97 (170) [0.31]	2.08 (170) [0.88]	4.76 (34) [0.63]	2.59 (34) [0.28]
2	-0.83 (177) [0.40]	2.01 (177) [0.66]	-3.40 (32) [0.51]	2.72 (32) [0.41]
3	3.23 (177) [1.29]	4.95 (177) [1.69]	-8.04 (31) [0.91]	3.98 (31) [0.36]
4	2.65 (116) [0.84]	0.31 (116) [0.08]	27.92 (15) [2.75]	22.67 (15) [1.84]
Overall	0.62 (640) [0.45]	2.03 (640) [1.43]	1.07 (112) [0.26]	0.25 (112) [0.05]

Table 8: ITT Effect

Table 8 presents results from an ITT analysis (so compliance behavior is ignored) broken down by grade and school classification (low/high), therefore the effects represent the gains in

test scores attributable to winning a scholarship. Non-bracketed numbers are treatment effect estimates for the appropriate subgroups. Numbers in parentheses are sample sizes. Bracketed numbers are the absolute value of treatment effect t-statistics for a null hypothesis of no treatment effect. Overall, across grades, there appear to be mostly positive effects of a scholarship offer. The only effect that is statistically significant at less than a .05 significance level, however, is for reading scores for children applying in the fourth grade from high-score schools. The corresponding math scores are not quite significant but of a similar direction and near significance. These effects seem quite extreme and certainly not terribly plausible. Their direction can be explained by two facts: (1) for the subset of children with observed pre-test scores and post-test scores there is a positive treatment effect of 16.57, and (2) for the subset of children in this subgroup for whom we observed pre-test scores, the children for whom we don't observed post-test scores had higher pre-test scores than the children for whom we do observe post-test scores (this information was incorporated into the non-response weights). What is particularly noteworthy, however, is that the t-statistic is so large. This points to a problem with using non-response weighting adjustments and complete cases to address such missing data problems: they cannot always reflect our uncertainty about the structure of this missing data particularly when the sample size is as small as it is for this subgroup.

Moreover, the sample sizes for all four subgroups of children applying from schools with high average test scores are quite small and these effects must all be regarded with caution. This sample size issue only worsens in the subsequent two analyses for which the effective sample sizes become even smaller.

The results in Table 9 were obtained from an analysis in which the treatment was defined as "program participation." That is, a child could only be labeled as having received the treatment if he won a scholarship (which entitled him also to receiving help in finding an appropriate school). Children who did not win scholarships but attended private school were still considered to be not receiving the treatment (therefore they are compliers, not always takers). These effects are similar to those in the preceding table although they are, in general, of larger magnitude. The t-statistics change only incrementally however. In this table the numbers in parentheses represent "effective sample sizes". These numbers correspond to the expected number of compliers for each subgroup; for this treatment definition these numbers simply subtract the expected

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	-1.33 (124.8) [0.31]	2.86 (124.8) [0.88]	6.71 (26.3) [0.64]	3.65 (26.3) [0.28]
2	-1.05 (139.4) [0.40]	2.52 (139.4) [0.66]	-4.43 (27.4) [0.50]	3.54 (27.4) [0.41]
3	3.86 (138.0) [1.28]	5.91 (138.0) [1.67]	-13.81 (17.7) [0.93]	6.83 (17.7) [.35]
4	3.03 (92.2) [0.84]	0.36 (92.2) [0.08]	27.12 (9.0) [2.75]	22.01 (9.0) [1.79]
Overall	0.77 (494.4) [0.45]	2.52 (494.4) [1.42]	1.53 (80.4) [0.26]	0.36 (80.4) [0.05]

Table 9: Effect of Program Participation

number of never takers from the sample sizes in Table 8. Predictably, these numbers are even smaller than before and hence the corresponding results are even less reliable.

The results in Table 10 represent the treatment effects for compliers in an analysis which defines the treatment as attendance at a private school. This treatment definition allows for compliers, never takers and always takers. Once again the magnitude of the effects increases in the vast majority of cases, however, the t-statistics across subgroups are only slightly altered, if at all. The effective sample sizes are (in most cases) even smaller for this analysis, with numbers as low as 7.8 for children applying in the 4th grade from schools with high scores.

In sum, the results from these analyses do not provide strong evidence in either direction, though there does seem to be some evidence for positive effects on test scores for a few older subgroups.

9 Notation for our Data Template

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

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	-1.55 (111.7) [0.31]	3.31 (111.7) [0.88]	6.71 (26.3) [0.63]	3.65 (26.3) [0.28]
2	-1.18 (124.2) [0.40]	2.85 (124.2) [0.66]	-5.11 (24.8) [0.51]	4.09 (24.8) [0.41]
3	4.41 (123.8) [1.26]	6.76 (123.8) [1.63]	-16.40 (16.1) [0.91]	8.12 (16.1) [0.36]
4	3.49 (84.3) [0.86]	0.41 (84.3) [0.08]	29.59 (7.8) [2.58]	24.02 (7.8) [1.89]
Overall	0.88 (444.0) [0.45]	2.89 (444.0) [1.42]	1.66 (75.0) [0.26]	0.39 (75.0) [0.05]

Table 10: Effect of Private School Attendance

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 SCSF program.

Deviations from the ideal scenario that occur frequently and are present in the SCSF program 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. 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. Our model will allow us to return to the scenario consisting of (1)–(4).

We now introduce the notation required for the formalization of the probability model corresponding to this 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[11], Y_i^{po}, which is composed of two P-length vectors, $Y_i(0)$ and $Y_i(1)$, where

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

[11] 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.

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 each has two potential values: one corresponding to each of the treatment assignments. $Y(0)$ and $Y(1)$ are used to denote the two $n \times p$ matrices of potential outcomes corresponding to control and treatment assignment respectively.

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 $n \times P$ 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. We can also think of R_{y_i} as the *intended* portion of $R_{y_i}(0)$ and $R_{y_i}(1)$.

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 relative test scores of the school the child attended at time of application (low/high), 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.

10 Pattern Mixture Model

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, indicator variables for 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, and not indicators for its missingness, must become part of the model. Fortunately in our setting, the major policy relevant covariates (grade, school test scores), 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 it is difficult to conceive of a new program targeting subgroups defined by these variables (e.g., it's difficult to imagine that pre-test scores would used as eligibility criteria for a program, whereas school test scores, perhaps as a proxy for school quality, could be used).

In Bayesian approaches, pattern mixture models typically factor the joint distribution of indicators for missing data patterns and data as the marginal distribution of the missing data patterns and the conditional distribution of the data given these patterns. The parameters of the conditional data model are typically under-identified; assumptions regarding missing data mechanisms can help to identify these parameters.

A variety of authors use pattern mixture model approaches to missing data including Rubin (1977, 1978b); Little (1993); Glynn, Laird, and Rubin (1986, 1993); Little (1996). Standard pattern mixture models partition the data with respect to the missingness of the primary vari-

ables of interest. In this application we partition the data with respect to only covariate missing data patterns, so the assumptions will differ slightly from the standard usage. One argument that can be made for the pattern mixture approach used in this setting is that it focuses the model on the primary quantities of interest, (functions of) Y or $Y \mid X$. The marginal distribution of X and R_X is ignored.

We describe the model first by stating 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.

10.1 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.

10.1.1 SUTVA

A standard assumption made in causal analyses of this kind is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978a, 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 did not 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, or, rather, we expect our results to be rather robust to the types and degree of deviations from SUTVA that we might expect in this study.

If we define the treatment as private school attendance, the no "versions of treatments" part of SUTVA is satisfied if the definitions of private school and public school encompass all the varieties of such schools encountered by the study participants. Similarly, if we define the treatment as participation in the scholarship program (winning the money, receiving help in finding a school), treatment homogeneity is satisfied or not depending on how rigidly "participation in the scholarship program" is defined – e.g. is using the money sufficient, or need the families all have received help in finding a new school as well?

10.1.2 Randomization

We assume scholarships have been randomly assigned. This implies

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

where W^* represents the portion of W that comprises the design variables, and θ is generic notation representing the parameters in any model. 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 test score classification (low/high) and application wave are known.

10.1.3 Missing data process assumption – Latent Ignorability

We assume that missingness is ignorable given observed covariates within subgroups defined by compliance status. Here, observed covariates includes indicators for missingness of the covariates, R_x, as well. This assumption 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. Note that this is a *non-ignorable* missing data mechanism.

Recall that latent ignorability differs from standard ignorability (Rubin 1978a; 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.

10.1.4 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, which generalized the standard exclusion restriction (Angrist, Imbens, and Rubin 1996; Imbens and Rubin 1997), 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.

Compound exclusion seems more plausible for never takers than for always takers. Never takers stay in the public school system no matter whether they win a scholarship or not. Always takers, on the other hand, might be in one private school if they won a scholarship or another if they didn't win a scholarship, particularly since those who won scholarship have access to resources to help find an appropriate private school. In addition, the scholarship provides the family with $1400 more in resources than is available to the family who didn't win a scholarship and still sends a child to private school; this could in and of itself have an effect on student outcomes.

10.1.5 Noncompliance process assumption II – Monotonicity

Implicit in the definition of compliance status, C, and as pointed out in Section 9, 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 (see, for example, Angrist, Imbens, and Rubin 1996) 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, $D_i(Z_i = 1) \geq D_i(Z_i = 0)$ (Imbens and Angrist 1994). In the SCSF program defiers would be families who would not use a scholarship if they won one, but, would

pay to go to private school if they did not win a scholarship. It seems highly implausible that such a group of people exists, therefore the monotonicity assumption appears to be quite reasonable.

10.2 Parametric Model

Our full model needs simultaneously to (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.

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

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

where $\theta = (\theta^{(\text{C})}, \theta^{(\text{R})}, \theta^{(\text{Y})})'$, 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.

10.2.1 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:

$$
\begin{aligned}
&C_i(n) = 1 \text{ if } C_i(n)^* \equiv g_1(W_i, X_i^{\text{obs}}, R_{x_i})' \beta^{(\text{C},1)} + V_i \leq 0 \\
&C_i(c) = 1 \text{ if } C_i(n)^* > 0 \text{ and } C_i(c)^* \equiv g_0(W_i, X_i^{\text{obs}}, R_{x_i})' \beta^{(\text{C},2)} + U_i \leq 0,
\end{aligned}
$$

where

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

The specific models attempt to strike a balance between including all the design variables as well as 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 11 use a compliance component model whose link function, g_1, fits, in addition to an intercept: school test scores (low/high); indicators for application wave; propensity scores for subjects applying in the first period and propensity scores for the other waves; indicators for grade of the student; recorded ethnicity (African American or other); 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. The link function g_0 is the same as g_1 with the exception that it excluded the indicators for application wave. This link function, a more parsimonious version of one we employed in earlier models, was more appropriate to fit the relatively small proportion of always-takers.

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

$$\beta^{(\mathrm{C},1)} \sim \mathrm{N}(\beta_0^{(\mathrm{C},1)}, \{\sigma^{(\mathrm{C},1)}\}^2\mathbf{I}),$$
$$\text{and} \quad \beta^{(\mathrm{C},2)} \sim \mathrm{N}(0, \{\sigma^{(\mathrm{C},2)}\}^2\mathbf{I}),$$

where $(\sigma^{(\mathrm{C},1)})^2$ and $(\sigma^{(\mathrm{C},2)})^2$ are "known" hyperparameters set at ten, and $\beta_0^{(\mathrm{C},1)}$ is a vector of zeros with the exception of the first element which is set equal to $-\Phi^{-1}(1/3)*\{1+\sigma^{(\mathrm{C},1)}\ \mathrm{ave}(g'_{1,i}g_{1,i})\}^{\frac{1}{2}}$, where $g_{1,i} = g_1(W_i, X_i^{\mathrm{obs}}, R_i)$, and ave denotes the average over the students. These priors reflect our a priori ignorance about the probability any individual belongs to each compliance status by setting each of their prior probabilities at 1/3.

10.2.2 Outcome Sub-Model

The specification for the outcome sub-model first posits a latent variable such that

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

for $z = 0, 1$, where $\theta^{(Y)} = (\beta^{(Y)}, \zeta^{(Y)})$ 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 1978a). Then

$$Y_i(z) = \begin{cases} 0 & \text{if } Y_i(z)^* \leq 0, \\ 100 & \text{if } Y_i(z)^* \geq 100, \\ Y_i(z)^* & \text{otherwise.} \end{cases}$$

The results reported in Section 11 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; school test scores (low/high); recorded ethnicity; 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; school test scores (low/high); recorded ethnicity; indicators for grade; the propensity score; 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. An indicator for whether or not a person is an always-taker.
4. An indicator for whether or not a person is a complier.
5. For compliers assigned treatment: an intercept, one indicator for school test scores (low/high); ethnicity; and indicators for 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:

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

and where $\zeta^{(\mathrm{Y})} = (\zeta_1, \ldots, \zeta_K)$, 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 five; and

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

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 (set at three) and scale parameter σ^2 (set at 400).

10.2.3 Outcome Response Sub-Model

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

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

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

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

The link function of the probit model on the outcome response, g_2, is the same as the link function for the mean of the outcome component.

The prior distribution for the outcome response sub-model is

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

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

11 Results

All of the results below were obtained from the same Bayesian analyses (one for math scores and one for reading scores). Both analyses include latent variables for compliers, never takers

and always takers, imposing the exclusion restriction on never takers and always takers. The differences between the results for the first three tables reflect different ways of averaging over the results for these groups, as described in each subsection.

The results are reported by school test scores classification (low/high) and grade – our "policy-relevant" variables – averaging over the other characteristics in the model. Both school test scores classification and grade were thought to have possible interaction effects with treatment assignment. Most of the following estimates are not parameters of the model but functions of parameters, whose posterior distributions are induced by the posterior predictive distributions (multiple imputation) of the compliance statuses. Except when otherwise stated, plain numbers are posterior means and brackets are 95% posterior intervals.

11.1 Test Score Results

In this section we 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 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. The individual-level causal estimates have also been weighted so that the subgroup causal estimates correspond to the effects for all eligible children belonging to that subgroup who attended a screening session. The numbers in parentheses represent either the sample sizes or "effective sample sizes" corresponding to each subgroup, just as described in Section 8, though here the posterior means of the parameters reflecting probabilities for each compliance category were used as estimated probabilities when calculating expected values for each.

11.1.1 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. We calculate the ITT effect by averaging over the effect in all three compliance groups[12].

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	2.06 (244) [-1.69, 5.54]	4.89 (244) [1.70, 8.05]	1.31 (46) [-4.74, 6.80]	5.01 (46) [0.07, 9.81]
2	0.20 (244) [-2.85, 3.44]	1.10 (244) [-2.08, 4.21]	-0.71 (45) [-5.99, 4.75]	0.97 (45) [-4.44, 6.55]
3	0.46 (233) [-3.00, 4.13]	3.02 (233) [-0.85, 6.66]	-0.60 (40) [-6.21, 4.81]	2.49 (40) [-3.99, 8.08]
4	2.78 (171) [–1.16, 7.06]	2.65 (171) [-1.50, 6.81]	1.69 (27) [-4.10, 7.81]	2.15 (27) [-3.76, 7.85]
Overall	1.27 (892) [-0.80, 3.42]	2.94 (892) [0.71, 5.15]	0.32 (158) [-4.89, 4.96]	2.73 (158) [-2.01, 7.15]

Table 11: ITT Effect

These results indicate posterior distributions primarily (i.e. greater than 97.5%) to the right of zero for the treatment effect on mathematics scores for 1st graders, and overall grades from low-score schools. Each indicate an average gain of more than 2.9 percentile points for children who won a scholarship.

11.1.2 Effect of participation in SCSF program

The results displayed in Table 12 reflect the effect of participation in the SCSF. They were calculated by measuring the ITT effect for always takers and compliers combined[13]. This analysis

[12]This strategy is an approximation to the most appropriate analysis for this estimand which would relax the exclusion restriction on both the always takers and never takers.

[13]This strategy is an approximation to the most appropriate analysis for this estimand which would relax the exclusion restriction on the always takers.

defines the SCSF *program* 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 SCSF program.

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	2.74 (216.9) [-2.19, 7.32]	6.40 (216.9) [2.24, 10.79]	1.76 (42.6) [-6.46, 9.39]	6.61 (42.6) [0.08, 13.00]
2	0.24 (214.7) [-3.50, 4.20]	1.38 (214.7) [-2.59, 5.29]	-0.97 (41.0) [-8.03, 6.12]	1.31 (41.0) [-5.76, 8.52]
3	0.60 (204.6) [-3.84, 5.18]	3.96 (204.6) [-1.08, 8.91]	-0.84 (36.4) [-8.37, 6.82]	3.46 (36.4) [-5.26, 11.42]
4	3.40 (152.7) [-1.44, 8.34]	3.23 (152.7) [-1.79, 8.34]	2.31 (24.9) [-5.65, 10.29]	2.93 (24.9) [-5.00, 10.20]
Overall	1.63 (788.9) [-1.06, 4.46]	3.74 (788.9) [0.88, 6.56]	0.43 (144.9) [-6.40, 7.11]	3.68 (144.9) [-2.61, 9.74]

Table 12: Effect of Scholarship Program

We see a similar pattern of effects in this analysis though the posterior means are all larger in absolute value than in the ITT analysis. The intervals are also larger than the ITT intervals which is not surprising given that the estimand now applies to only a subset of the study participants (as reflected by the effective sample sizes in parentheses).

11.1.3 Effect of Private School Attendance

The results in Table 13 represent the effect of private school attendance by focusing only on the compliers. This analysis defines the "treatment" as private school attendance. The validity of these results rest, in part, 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.

Grade at Application	Low		High	
	Reading	Math	Reading	Math
1	3.08 (164.2) [-2.56, 8.38]	7.24 (164.2) [2.45, 12.41]	1.89 (32.2) [-7.2, 10.53]	7.14 (32.2) [0.09, 14.36]
2	0.28 (171.8) [-3.94, 4.80]	1.57 (171.8) [-2.98, 6.05]	-1.08 (30.9) [-9.49, 6.75]	1.44 (30.9) [-6.10, 9.68]
3	0.67 (155.9) [-4.28, 5.81]	4.53 (155.9) [-1.22, 9.50]	-0.93 (26.2) [-9.33, 7.25]	3.82 (26.2) [-5.61, 12.58]
4	3.84 (125.2) [-1.76, 9.83]	3.62 (125.2) [-1.95, 9.50]	2.55 (18.4) [-6.15, 11.54]	3.17 (18.4) [-5.29, 11.12]
Overall	1.85 (617.1) [-1.21, 5.13]	4.24 (600.6) [0.99, 7.44]	0.47 (107.7) [-7.20, 7.77]	4.02 (107.7) [-2.83, 10.79]

Table 13: Effect of Private School Attendance

The effects of private school attendance, displayed in Table 13 are quite similar to the scholarship program effects with posterior means that are slightly bigger in absolute value than in the other two analyses. The intervals have also grown reflecting the still smaller effective sample sizes. The effective sample sizes for the subgroup of 4th-graders applying from high-score schools is so small (24.9) as to make these results a bit suspect.

11.2 Composition of compliance status

Table 14 gives estimates of the composition of compliance status as a function of school test scores classification and grade. Because the distributions between the two models (mathematics/reading) were comparable in both location and uncertainty (see also Section 11.3.1 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, in most cases, high-score schools have more never takers, fewer always takers, and slightly fewer compliers than low-score schools.

Grade at Application	School Test Scores	Never Taker	Complier	Always Taker
1	High	24.6 (5.8)	69.9 (6.6)	5.5 (3.3)
	Low	24.2 (3.5)	67.3 (4.3)	8.5 (2.7)
2	High	24.7 (5.7)	68.7 (6.5)	6.6 (3.4)
	Low	20.0 (3.4)	70.4 (4.3)	9.6 (2.7)
3	High	28.0 (6.3)	65.6 (7.0)	6.4 (3.3)
	Low	23.8 (3.8)	66.9 (4.5)	9.3 (2.6)
4	High	26.3 (6.2)	68.0 (6.8)	5.7 (3.2)
	Low	18.0 (3.7)	73.2 (4.8)	8.8 (3.1)

Posterior standard deviations are in parentheses.

Table 14: Composition of compliance status

11.3 Impact of missing data

When the latent compliance groups have differential response (i.e. missing data) 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 11.3.1 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 are skewed, so posterior medians and posterior intervals are reported.

For each of the first two comparisons (columns three and four), 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 (right-most column), any differences

Grade at Application	School Pre-Treat. Scores	Control Complier vs. Never Taker	Treatment Complier vs. Always Taker	Treatment Complier vs. Control Complier
1	High	2.4 [1.2, 5.1]	0.1 [0.0, 2.7]	1.7[0.5, 6.2]
	Low	2.3 [1.2, 4.5]	0.1 [0.0, 0.8]	1.4[0.7, 2.9]
2	High	2.3 [1.2, 5.0]	0.3 [0.0, 4.5]	2.9[0.9, 14.3]
	Low	2.1 [1.1, 4.2]	0.2 [0.0, 1.5]	2.4[1.1, 5.4]
3	High	2.4 [1.2, 6.1]	0.1 [0.0, 1.9]	2.0[0.5, 11.1]
	Low	2.3 [1.2, 4.9]	0.1 [0.0, 1.1]	1.6 [0.7, 3.8]
4	High	2.0 [1.0, 4.2]	0.1 [0.0, 1.7]	2.0 [0.5, 11.1]
	Low	2.1 [1.1, 4.1]	0.1 [0.0, 0.8]	1.6 [0.7, 3.9]

Results are reported combined from mathematics and reading models because they were similar

(raw data on response in mathematics and reading were identical).

Numbers are posterior medians and posterior 95% intervals.

Table 15: Odds ratios comparing response rates among groups.

are attributed to the treatment. From the table it can be deduced that response is increasing in the following order: never-takers, compliers attending public, compliers attending private, and always-takers. Therefore, the latent compliance behavior appears to be an important predictor of response.

11.3.1 Agreement between the models

Before running the final analyses, we assessed the agreement between the model for mathematics and the model for reading in predicting compliance type (i) at the individual student level, and (ii) as a function of the covariates low/high 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 whose compliance type was not known, we computed 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 low/high the agreement of the posterior distributions of compliance status, summarized by posterior first two moments, was very good.

12 Comparison Between Models

The analyses relying on standard approaches presented in Section 8 (henceforth referred to as MPR analyses) and the Bayesian analyses were performed on the same outcomes and for the same initial subset[14] of children (single-child families from grades one through four). In addition they both attempt to address the same complications which draw the template away from the perfectly controlled randomized experiment. The Bayesian analyses, however, rely on weaker structural (though perhaps slightly stronger parametric) assumptions than the MPR analyses. These strategies, therefore, invite comparison.

Results from the Bayesian analyses lead to somewhat similar, although not altogether consistent, inferences to those indicated by the MPR analyses, largely in the sense that neither analysis shows consistently strong evidence in one direction or another. If we examine the overall results, there is a fair amount of agreement between the approaches (for all three questions asked) for math scores of children applying from low-score schools, across all grades. Both approaches provide evidence for positive gains of two to three and three-quarters percentiles for math scores of children from low-score schools.

There does appear to be a difference with regard to the effect on the other scores. For the effect on reading scores of children who applied from low-score schools, the Bayesian analyses report generally positive, though not very strong, gains across grades, whereas the MPR analyses show both positive and negative mean effects. Another, more specific difference exists on reading and mathematics on 4th graders from high-score schools, between the modest effects reported by the Bayesian analyses and the large effects reported by MPR analyses. Section 8

[14]Clearly the exclusion of students with missing pre- and post-test scores as a part of the approach to handling these missing data problems creates a non-randomly smaller sample for the MPR analyses. However both intend their inferences to apply to the same population.

briefly discusses this difference in terms of the issues involved with non-response weighting. In general, these differences are driven by the fact that these analyses condition on different sources of information. The Bayesian analyses can include subjects with missing pre-test scores or post-test scores.

One way in which these different sources of information affect inferences concerns the differences between children who took the pre-test and those who didn't. As it turns out, for the people with missing pre-test scores (excluded in the MPR analysis), there is a negative treatment effect for reading post-test scores, which serves to counteract, in part, the positive treatment effects we see in those for whom we observe pre-test scores. Another difference has to do with the amount of smoothing allowed in each model. Currently our Bayesian models are quite parsimonious, so it is possible that we haven't allowed for enough cell to cell variation in treatment effects. The difference between the two model approaches within subgroups defined by grade and type of school is also likely influenced by the relatively smaller sample sizes in these subgroups. Clearly the larger the sample size, the greater chance we have of finding consistent results across the two approaches.

13 Discussion

Future analyses will attempt to learn from these initial models and incorporate additional complexity. We would like to investigate more closely the appropriateness of additivity and smoothing in the model, include both math and reading outcomes in the same model, include more covariates, and test the sensitivity to relaxing each of the exclusion restrictions. With additional years' data we will have to model appropriately the time series nature of the data as well as the more complicated compliance structures that will develop. We also recognize the need to perform more model checks and sensitivity analysis than we have performed to date.

As far as substantive conclusions regarding school choice, both models appear to indicate gains in math scores for children from low-score schools who have either won a scholarship, participated in the scholarship program or attended private school; however, neither analysis strategy leads to convincing and consistent overall evidence in favor of private schools. If it is truly the case the we see greater gains for the children from low-score schools, this information

would have policy implications and would provide greater justification for the current Florida school choice initiative which targets more disadvantaged schools. Given the heterogeneity in schools and the noisiness of our outcome measure, it is likely that it is too soon to expect to find sharp differences between the treatment and control groups.

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.

We would also like to acknowledge support for the methodological work from the National Institute of Child Health and Human Development (R01 HD38209), National Institute of Mental Health (R01 MH56639), National Institute for Drug Abuse (R01 DA10184), the H.-C. Yang Memorial Faculty Fund, and National Science Foundation (SBR 9709359 and DMS 9705158).

The introduction to this paper 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, Rubin, and Thomas (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.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986), "Mixture Modeling Versus Selection Modeling for Nonignorable Nonresponse," in *Drawing Inferences from Self-Selected Samples*, ed. H. Wainer, pp. 115–142, Springer-Verlag.

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

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

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., Rubin, D. B., and Thomas, N. (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," to appear in *Biostatistics.*

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

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

Imbens, G. W. and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *The Annals of Statistics* 25, 305–327.

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

Little, R. J. A. (1993), "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association* 88, 125–134.

Little, R. J. A. (1996), "Pattern-mixture models for multivariate incomplete data with covariates," *Biometrics* 52, 98–111.

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

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.

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.

Peterson, P. E., Myers, D. E., Howell, W. G., and Mayer, D. P. (1999), "The Effects of School Choice in New York City," in *Earning and Learning; How Schools Matter*, eds. S. E. Mayer and P. E. Peterson, Brookings Institution Press.

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

Rubin, D. B. (1978b), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse (C/R: P29-34)," in *ASA Proceedings of Survey Research Methods Section*, pp. 20– 28.

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.

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 possibly latent variables $Y_i^* \equiv Y_i(Z_i)^*$ for the outcome model; all 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 with conditioning that ensures convergence of the Gibbs Sampler to the posterior distribution. 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 distribution. Moreover, at this step, the conditioning on Y_i^* and $R_{yi}(Z_i)^*$ can be replaced, respectively, by $Y_i * R_{yi}(Z_i)$ and $R_{yi}(Z_i)$ for algorithmic efficiency. 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 $p(C_i, Y_i, D_i, R_{yi}|Z_i, H_i, \phi)$. 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

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

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 Y_i^* is from $p(Y_i^*|H_i, C_i, Y_i, z = Z_i, \phi)$: when Y_i is in $(0, 100)$, $Y_i^* = Y_i$; when $Y_i = 0$ then Y_i^* is drawn from the tail of the defining normal distribution $p(Y_i^*|H_i, C_i, z = Z_i, \phi)$ left of 0, and the method of simulation is as with the compliance latent normals; there was no observation equal to 100.

4. 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 and outcome latent normals.

5. 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 the distribution $p(\beta^{(c,2)}|\{$ 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.

6. 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, $\widetilde{Y}_i = |Y_i^* - g_2(H_i, C_i, Z_i)\beta^{(Y)}|$ for the observed outcomes we have

that

$$\frac{(\widetilde{Y}_i)^2}{\exp(g_3(H_i, C_i, Z_i)\zeta^{(Y)})} \sim \chi^2_1, \quad \text{so}$$
$$E(2\log(\widetilde{Y}_i)) = E(\log(\chi^2_1)) + g_2(H_i, C_i, Z_i)\zeta^{(Y)},$$

where χ^2_1 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 on all parameters including $\zeta^{(Y)}$. Using the regression estimates from the last relation, we obtained the two moments for 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 5000 iterations. We assessed convergence with add hoc methods. Then a main series of an additional 5000 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
Low/high-score 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)	
Pre-test reading score (normal curve equivalent)	

continued on next page

continued from previous page	
Variable	Description
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)	

Appendix C – Finer Strata.

The model of Section 10.2 can be used to estimate the effect of the program on finer strata that may be of interest. For example, to estimate effects stratified by ethnicity, models for compliance, outcomes, and response analogous to those in Section 10.2 were estimated with parameters for recorded ethnicity (dichotomized as African American (AA) or other). As in Section 11, the models induce a posterior distribution for the causal effect of the program for each child, and children are subsequently stratified by the variables: grade (1-4), type of child's originating school's past scores (high/low), and ethnicity. The results for this stratification are reported in Tables 17, 18, 19, for the estimands of ITT, effect of scholarship, and effect of attendance of private versus public school respectively.

The results, as in those of Section 11, support evidence for effect on mathematics for certain subgroups of children. Here, the effects on mathematics are largest for African American children (with tighter intervals around the estimates in the low-score schools which constitute on average 85% of this ethnic sub-sample), smallest for non African Americans originating from high past-score schools, and in the middle range for the remaining students.

Grade at Application	Ethnicity	Low		High	
		Reading	Math	Reading	Math
1	AA	2.75 [-1.65, 6.58]	6.40 [2.59, 10.12]	1.84 [-4.93, 7.77]	6.25 [0.70, 12.03]
	other	1.43 [-2.36, 5.53]	3.56 [0.25, 7.21]	0.60 [-4.95, 5.89]	3.30 [-1.77, 8.10]
2	AA	0.72 [-2.89, 4.50]	2.28 [-1.20, 6.14]	-0.25 [-6.77, 6.20]	2.22 [-3.88, 8.38]
	other	-0.21 [-3.22, 3.23]	0.18 [-3.31, 3.77]	-1.09 [-6.16, 4.03]	-0.17 [-5.51, 4.99]
3	AA	0.99 [-2.94, 4.90]	4.43 [0.03, 8.58]	-0.06 [-6.72, 6.46]	4.29 [-3.10, 11.02]
	other	-0.03 [-3.55, 3.78]	1.73 [-2.40, 5.56]	-0.91 [-6.13, 4.14]	1.33 [-4.40, 6.49]
4	AA	3.54 [-0.94, 8.44]	4.03 [-0.54, 9.03]	2.45 [-4.10, 10.06]	3.66 [-2.88, 10.37]
	other	2.05 [-2.09, 6.24]	1.29 [-3.06, 5.83]	1.12 [-4.64, 6.48]	0.88 [-5.07, 6.46]

Table 17: ITT Effect.

Grade at Application	Ethnicity	Low		High	
		Reading	Math	Reading	Math
1	AA	3.36 [-1.98, 7.89]	7.70 [3.18, 12.28]	2.27 [-6.14, 10.09]	7.62 [0.86, 14.71]
	other	2.06 [-3.34, 7.92]	5.02 [0.35, 10.08]	0.92 [-7.53, 9.09]	5.04 [-2.51, 11.88]
2	AA	0.82 [-3.23, 5.06]	2.61 [-1.39, 7.00]	-0.32 [-8.27, 7.42]	2.70 [-4.56, 10.35]
	other	-0.29 [-4.33, 4.30]	0.25 [-4.43, 5.06]	-1.67 [-9.03, 6.04]	-0.24 [-7.73, 7.64]
3	AA	1.20 [-3.69, 5.85]	5.30 [0.03, 10.36]	-0.07 [-8.25, 7.92]	5.19 [-3.75, 13.12]
	other	-0.06 [-5.11, 5.40]	2.51 [-3.34, 8.22]	-1.44 [-9.60, 6.32]	2.11 [-6.68, 10.06]
4	AA	4.00 [-1.08, 9.42]	4.53 [-0.60, 10.15]	2.90 [-4.96, 11.49]	4.29 [-3.45, 12.21]
	other	2.72 [-2.74, 7.98]	1.71 [-3.86, 7.57]	1.68 [-7.12, 9.85]	1.32 [-7.41, 9.22]

Table 18: Effect of Scholarship Program.

Grade at Application	Ethnicity	Low		High	
		Reading	Math	Reading	Math
1	AA	3.75 [-2.15, 8.91]	8.66 [3.52, 13.78]	2.44 [-6.83, 10.87]	8.18 [0.93, 15.46]
	other	2.34 [-3.71, 9.59]	5.72 [0.42, 11.79]	1.00 [-8.37, 10.02]	5.51 [-2.66, 12.70]
2	AA	0.92 [-3.59, 5.68]	2.91 [-1.55, 7.88]	-0.35 [-9.03, 9.02]	2.92 [-4.90, 11.17]
	other	-0.34 [-5.10, 4.94]	0.29 [-5.03, 5.70]	-1.90 [-10.78, 6.73]	-0.23 [-8.25, 8.69]
3	AA	1.34 [-3.96, 6.72]	6.00 [0.04, 11.87]	-0.07 [-8.76, 8.73]	5.61 [-3.86, 14.67]
	other	-0.09 [-6.47, 6.30]	2.91 [-3.96, 9.73]	-1.65 [-10.83, 7.82]	2.39 [-7.32, 11.67]
4	AA	4.44 [-1.20, 10.91]	5.01 [-0.67, 11.37]	3.13 [-5.43, 12.31]	4.55 [-3.69, 13.03]
	other	3.12 [-3.13, 9.10]	1.96 [-4.46, 8.70]	1.88 [-7.95, 10.78]	1.46 [-8.05, 10.46]

Table 19: Effect of Private School Attendance.