STUDYING THE CAUSAL EFFECTS OF INSTRUCTION WITH APPLICATION TO PRIMARY-SCHOOL MATHEMATICS

Stephen W. Raudenbush

Guanglei Hong

Brian Rowan

University of Michigan

November 11, 2003

Prepared for

Research Seminar II
Instructional and Performance Consequences of High-poverty Schooling

March 11, 2002
The Charles Sumner School Museum and Archives
Washington, D.C.

Sponsored by

Office of Educational Research and Improvement
National Center for Educational Statistics
National Institute on the Education of At-Risk Students
Planning and Evaluation Service
American Institutes for Research

This research was supported by Westat subcontract 9844 – Longitudinal Evaluation of School Change and Performance (LESCP): A Secondary Analysis. The statistical algorithms used here were developed with support from The Project on Human Development in Chicago Neighborhoods with funding from the John D. and Catherine T. MacArthur Foundation and the National Institute of Justice. Richard Congdon deserves special thanks for the applications programming. We wish to thank Susan Murphy for her numerous and extremely helpful discussions about the issues we address in this paper.

DRAFT: Do not cite without the authors' permission.

STUDYING THE CAUSAL EFFECTS OF INSTRUCTION WITH APPLICATION TO PRIMARY-SCHOOL MATHEMATICS

Introduction

Every aspect of schooling in the United States is currently under intense scrutiny. Should the federal government demand high standards for student achievement? Should districts dole out cash rewards to schools and teachers performing well? Can private market mechanisms like vouchers and school choice improve student learning? As the public debates these questions about school governance, it also considers new initiatives to vastly increase resources for schooling: smaller classes, better qualified teachers, links to the internet in every classroom.

Corresponding to these intense debates over school reform is new interest in making educational research more scientific. Recent randomized experiments on vouchers (Peterson & Hassel, 1998), class size reduction (Finn and Achilles, 1990; 1999), and whole-school reform (Cook, Habib, Philips, Settersten, Shagle, and Demirmencioglu, 1999; Cook, Murphy, and Hunt, 2000) express a new determination to increase rigor in educational research. The idea is to inject authoritative evidence into the highly politicized debate over educational reform.

Parallels with the movement for clinical trials in medicine in the 1950s and 1960s are unmistakable and not coincidental. Operating independently, senior scholars (Boruch and Mosteller, 2002) and federal policy-makers (U.S. Department of Education, February, 2002) forcefully advocate that a new emphasis on experiments can do for education what clinical trials have done for medicine: to submit claims about good practice to definitive tests and thereby constrain the role of partisan politics and personal interest in decision making.

Making the Medical Analogy Work

There is a mismatch, however, in the analogy between the movement for clinical trials 50 years ago and the current focus of most policy research in education. At the heart of medical research are clinical trials that assess the causal effects of alternative clinical practices: medical procedures, surgical procedures, and vaccines. The question for health services research is then how to organize the delivery of good clinical practice in cost effective and equitable ways.

In contrast, it is currently possible for debate about school governance, school management, and educational resources to float free of any reference to the heart of educational practice: classroom instruction. We know too little about the causal effects of alternative instructional strategies. As a result, our research on school management, governance, and resources can't tell us what we need to know to improve schooling.

Cohen, Raudenbush, and Ball (2002) review four decades of research on educational resources. While producing many intriguing findings, this research bears weakly on practice, and the overall scientific yield has been disappointing. The essential problem is that classroom instruction -- the essential clinical practice of education -- is absent from the causal models that have driven most research on resources.

The logic in medicine seems to work as follows: clinical practices are the proximal causes of patient outcomes. Clinical trials establish knowledge about the causal effects of these practices. Health services research considers how to deliver those practices equitably and cost effectively.

The parallel in education ought to be clear. Instructional practice in the classroom is the proximal cause of students' academic learning. Evaluation of the causal effects of instruction would supply the foundation of knowledge for educational improvement. In this scheme, research on whole-school reform or accountability or governance would then consider cost effective ways of supporting effective instruction.

Firm knowledge about practices that reliably produce skill and knowledge in mathematics, for example, would clarify the skill requirements for primary school teachers of mathematics, shaping, therefore, new research questions about inservice and preservice education. It would pose clear questions about class size. It would clarify the sequences of experiences children need as they traverse the primary school grades and therefore suggest how schools might best coordinate and support instruction, grounding research on whole school reform in the realities of teaching and learning specific subject matter.

In sum, a commitment to learn about how alternative medications and surgical procedures affect patient health is the bedrock of advances in medical research over the past 20 years. If educational research is to emulate the medical model, learning about instructional interventions has to be the foundation of the new movement to make educational research more scientific.

The Logic of Instructional Research

Educational researchers have paid considerable attention to uncovering promising new ideas about instruction. They have meticulously observed expert teachers in action, described the diversity of ways children learn in classrooms, and, based on these studies, proposed coherent new standards for math and other subjects and new models for instruction (Grouws, 1992; Pellegrino, Baxter, and Glaser, 1999; Richardson, 2001; Wittrock, 1986).

Developing promising models provides an essential first step in generating useful knowledge for policy and practice. The next step is to test the efficacy of these alternative models across a reasonably diverse array of classrooms settings. This requires the explication of the models so that they can be replicated, at least approximately, by varied teachers and contrasted in a controlled way to plausible alternative models. There have been numerous successful attempts to do this, especially in primary school reading (Pinnell, DeFord, Lyons, and Bryk, 1995; Snow, Burns, and Griffin, 1998) and writing (Englert, 1991).

Unfortunately, causal-comparative studies of instruction constitute a small fraction of research on instruction. When they do occur, such studies tend to be bounded by a single academic year. The academic year provides a natural focus for studying instruction. And as we shall see, the methodological challenges to valid causal inferences about instruction are comparatively manageable when the study is confined to the single academic year.

While understanding instructional effects during a single year is necessary to understanding those effects more generally, it cannot be sufficient. For example, there is good reason to worry that a student who cannot decode familiar text by the end of grade 1 will not read new text with high comprehension in grade 3. But decoding in grade 1 does not insure comprehending in grade 3. To understand how primary school instruction affects comprehension by the end of grade 3 requires study of the causal effects of a *sequence* of instructional experiences occurring from kindergarten through the end of grade three.

Similarly, computational skill in addition in grade 1, even if grounded in a solid conceptual understanding of the number line and its simple applications, does not insure that, by grade 4, a student will be able to organize available data in a way that bears on solving a multi-step arithmetic problem. To understand primary school instruction requires, once again, an understanding of the causal effects of a sequence of experiences in math instruction unfolding over multiple years.

The Current Paper

In this paper, we consider the methodological challenges facing any ambitious new attempt to assess the causal effects of alternative instructional "treatments." Studies of instruction that are bounded by a single year follow a familiar logic: one searches for designs and analytic methods that control pre-treatment ("baseline") covariates related to the outcome that also predict treatment group membership. Randomized experimental designs are plausible in this scenario. When they are not possible, methods like propensity score matching (Rosenbaum and Rubin, 1983a) and linear model adjustment (Cochran, 1957) followed by sensitivity analysis (c.f., Rosenbaum, 1986; Rosenbaum & Rubin, 1983b) can often be helpful. Statistical methods are required that take into account the clustering of students within classes and schools, a familiar problem with now-standard solutions (c.f., Raudenbush and Bryk, 2002).

Studies of the effects of a sequence of instructional treatments pose far greater challenges. The key problem is that treatment-group membership varies with time and may be confounded with time-varying covariates related to earlier treatments as well as earlier outcomes and baseline covariates. Randomized experiments are much more difficult to implement and sustain in this setting. Nor do standard methods of adjustment such as propensity-score matching or linear-model adjustment work well.

We draw on a new line of research in epidemiology for methods to solve this problem. Robins and his colleagues (Robins, 1987, 2000; Robins, Herman, and Brumback, 2000) have intensively studied the problem of time-varying treatment exposure and have proposed satisfactory methods for taking into account time-varying as well as time-invariant covariates. We explicate how this approach can be applied to instruction, and adapt it to the complex multilevel setting that arises in longitudinal instructional research, in which repeated measures are cross-classified by students and teachers who are, in turn, nested within schools.

The paper is organized as follows. Section 2 considers methodological challenges that arise in studying instructional treatments. We consider how randomized experiments might be designed to study effects of alternative treatments implemented during a single year as well as alternative sequences of treatments over multiple years. We then discuss challenges that arise when randomized experiments are impossible or when they unravel over time. Section 3 of the paper introduces an illustrative example. We show how our general approach applies to an assessment of the causal effects during grades 4 and 5 of "intensive" mathematics instruction, that is, classroom instruction that emphasizes comparatively high-level content and devotes substantial classroom time to mathematics. Data are from the Longitudinal Evaluation of School Change and Performance (LESCP) collected as part of an evaluation of Title I. In this example, we show how to tailor a hierarchical linear model for these data, applying inverse-probability-of-treatment weighting to adjust for prior outcomes, treatments, and measured baseline and time-varying covariates. Section 4 provides the results of the illustrative analysis. Stability and sensitivity analyses lead us to conclude that intensive math instruction as we have defined it has practically significant positive effects in fifth but not fourth grade. Section 5 concludes by considering how the approach we illustrate here might be applied more broadly in instructional research and by highlighting some unsolved methodological challenges.

Methodological Challenges to Assessing the Causal Effects of Instruction

Understanding a sequence of instructional effects is not equivalent to understanding the effects of a sequence of instructional experiences. The effect of next year's experience may well depend on what happened to you this year. More formally, knowing the average causal effect of a new method of instruction A' versus an old method A in grade 1 and also knowing the average causal effect of method B' verus B in grade 2 does not tell us the average causal effect of the sequence (A',B') versus (A,B) across the two grades. The average benefit of experiencing A' versus A in grade 1 might be *a*, and the average benefit of B' versus B in grade 2 might be *b*. But it does not follow that the average benefit of A' and then B' is *a* + *b*. Indeed, if B' builds deliberately and effectively on A', the benefit of the sequence may substantially exceed the sum of these two average effects.

Assessing the causal effect of a sequence of instructional experiences is much more challenging than assessing the causal effect of instruction during a single year. Let us consider each of these scenarios in turn. In each case we consider causal inference under a randomized experiment versus causal inference without benefit of a randomization.

Estimating the Causal Effect of Instruction During One Academic Year in a Randomized Experiment

The best design for estimating causal effects, if it can be implemented successfully, is the randomized experiment. For example, we might randomly assign children in equal numbers to receive method A' versus method A of math instruction during grade 1.[1] At the end of grade 1, we would compute the average math proficiency (or the average gain) for the two groups. The difference between the two averages would be an unbiased estimate of the average causal effect of A' versus A. In the absence of random assignment, such a difference between group averages cannot be regarded as an unbiased estimate of the causal effect.

Statisticians have reached a near consensus about how to conceive such causal effects (Holland, 1986; Rosenbaum and Rubin 1983a; Rubin, 1974, 1977, 1978). Sometimes known as the "Rosenbaum-Rubin-Holland" theory of causal analysis, this approach is also referred to as the "potential outcomes" framework for causal analysis.

[1] To assume that treatment A versus A' is the true cause of any observed effect requires random assignment of classrooms or whole schools to A versus A'. Using clusters as units of randomization slightly complicates the analysis but follows the same basic logic as noted here. We treat the students as units of random assignment in this hypothetical scenario for simplicity of exposition. For simplicity we also assume for now that A and A' are implemented with equal fidelity by all teachers in those treatments and that students' responses to treatment do not depend on which other students are assigned to the treatment. We relax these unrealistic assumptions in the illustrative examples. These assumptions correspond to the "Stable Unit-Treatment Assumption" of Rubin (1978).

Under this conception, and under the assumptions described in Note 1, we conceive each child to have two potential outcomes in the context of our hypothetical experiment: the outcome that the child would display if assigned to method A' and the outcome that the same child would display if assigned to method A.[2] The difference between these potential outcomes is the causal effect of method A' relative to A. Note that the causal effect so defined is child-specific. However, this child-specific effect cannot be observed because the child will experience only one of the two instructional "treatments." If the child is assigned to method A' we will never know the "counterfactual outcome," that is, how the child would have done under A. Similarly, if the child is assigned to A, we will not observe the counterfactual outcome under A'. This is a missing data problem, which Holland (1986) describes as "the fundamental problem of causal inference." Although we cannot estimate child-specific causal effects, we can estimate average causal effects without bias, for the population as a whole or for sub-populations, if we design our studies well and do the analysis right.

A causal model for this analysis appears in Figure 1. The variable $Z = 1$ if a student is in A' while $Z = 0$ if that student is in A.[3] $Y = (Y^{(1)},Y^{(0)})$ is the pair of potential math outcomes for each child, where $Y^{(1)}$ is the outcome that will be observed if the child experiences treatment 1 and $Y^{(0)}$ is the outcome that will be observed if the child experiences treatment 0[4]. The fact that assigning $Z = 1$ as compared to $Z = 0$ would produce a difference $Y^{(1)} – Y^{(0)}$ in outcomes is indicated by the single-headed ("causal") arrow that goes from X to Y.

---

[2] A more satisfactory conceptual framework envisions a potential outcome for each instructional method and for each teacher that student might encounter. We stay with the simpler notion of two potential outcomes here and extend it later to include 2*J potential outcomes, where J is the number of teachers the student might encounter.

[3] Henceforth we shall refer to the two treatments as $Z = 1$ versus $Z = 0$ rather than A' versus A.

[4] As mentioned both potential outcomes will never be observed. If the student is assigned to $Z = 0$, we will see $Y^{(0)}$ but not $Y^{(1)}$. If assigned to $Z = 1$, we will see $Y^{(1)}$ but not $Y^{(0)}$.

Insert Figure 1 About Here

Also in the figure is the symbol "U" which denotes all unmeasured pre-treatment predictors of potential outcomes Y. ("Pretreatment" predictors of Y are those that could have been measured prior to assignment to grade 1 instructional treatments). Because students were randomly assigned to treatment conditions, we can assume that U is not associated with Z. We denote this independence symbolically as

$$Y^{(0)}, Y^{(1)} \perp Z \quad or \;\; U \perp Z \,. \qquad \textbf{(1)}$$

The symbol "⊥" denotes statistical independence. Equation (1) says that the potential outcomes ($Y^{(0)}$,$Y^{(1)}$) of the student are statistically independent of treatment assignment ($Z = 1$ or $Z = 0$). Equivalently, knowledge about U is of no use in predicting Z. This equation would be false, if, for example, the parents knew that their child would do better under $Z = 1$ than $Z = 0$ (that is, they knew $Y^{(1)} > Y^{(0)}$) and so selected $Z = 1$ for their child. Under random assignment, that kind of prescient choice is impossible. Equation 1 would also be false if a child with high average achievement (e.g., high $[Y^{(0)} + Y^{(1)}]/2$) had especially high odds of landing in one treatment or the other. Random assignment also forbids that kind of "selection bias."

In this setting, the causal effect for each child is the difference between potential outcomes, that is, $Y^{(1)}$ - $Y^{(0)}$, and the average causal effect is the average of these differences in the population of interest. We denote this average causal effect as $E(Y^{(1)} - Y^{(0)}) = \delta,$ where "E(.)" denotes an expected value.

Randomized experiments of this type have often occurred in education (c.f., Englert, 1991). However, in randomized experiments, students or teachers may drop out of the study in disturbingly large numbers or students may not appear on the day of the final test. Some teachers may decide not to comply with the instructional method to which they are assigned. In these circumstances, we lose the benefit of the randomization. The variables *U* that predict the potential outcomes might also cause teachers to drop out or students to leave the school or students to be absent on the day of the test. Then, suddenly, *U* becomes associated with *Z* and Equation 1 is false. When this equation is false, the mean difference between the two groups at the end of the study is no longer an unbiased estimate of the average treatment effect (c.f., Holland, 1986).

In other cases, it is not possible to randomly assign students (or classrooms or schools) to treatments. If a randomized experiment deteriorates because of missing data or non-compliance or if a random assignment was not possible, we cannot assume Equation 1. Then valid causal inference is more challenging.

Understanding the causal effect of Instruction During One Academic Year in a Non-randomized Study

Suppose now that students are assigned to treatment 1 versus 0 not by random assignment but as a result of a combination of parent choice, teacher choice, and a variety of other systematic or accidental factors. Then we can no longer assume that the potential outcomes are independent of treatment assignment.

When random assignment is not possible, social scientists take pains to measure a variety of *covariates*, that is, pre-treatment characteristics of students or teachers that might be related to the potential outcomes. Let the measured covariates be labeled *X*, as distinct from those that were not measured, which we will call *U*'.

The social scientist must worry about confounding variables or "confounders." A confounder is a pre-treatment predictor of the potential outcomes *Y* that also predicts treatment-group assignment, *Z*. While the social scientist would like to measure and control all pre-treatment predictors of *Y*, she will be quite content if she successfully identifies and measures all *confounders*. That is, she will be happy if the measured covariates, *X*, include all confounders, meaning that the unmeasured covariates, *U*', are not related to *Z* once *X* is taken into account. More concretely, if we sub-divide the population of students into strata, each of which has the same value of *X*, we are happy if differences in *U*' are independent of *Z within those strata*. Symbolically, we would write

$$Y^{(0)}, Y^{(1)} \perp Z|X \quad or \quad U' \perp Z|X \qquad \textbf{(2)}$$

which is to say "the potential outcomes ($Y^{(0)}$, $Y^{(1)}$) are independent of the treatment assignment (*Z*) *given* ("holding constant") *X* (the vertical bar is a conditional and is to be read "given"). Put another way, we could say that among children having the same value of *X*, there is no association between their potential outcomes and their assignment to treatment. After taking *X* into account, *U*' is of no use in predicting treatment assignment *Z*.

Equation 2 is known as the assumption of "strong ignorability" (Rosenbaum, 1984; Rosenbaum and Rubin, 1983a; Rubin, 1978): in evaluating the causal effects of *Z* we can ignore *U*' as long as we control for *X*. It is a "strong" assumption because it requires that we have been successful in measuring all confounders, that is, that all confounders are in *X* and there are none in *U*'.

When strong ignorability holds, we can obtain an unbiased estimate of the average causal effect of the treatment as follows: compute the mean difference between the two treatments within every level of *X*; then average these.[5] Note that we cannot compute treatment effects within levels of *X* unless we find

[5] One would use a weighted average because the strata will provide different amounts of information about the causal effect.

within those levels some children who did experience $Z = 0$ and others who experienced $Z = 1$. If levels of X are completely segregated by Z, so that X and Z are perfectly confounded, it is not possible to estimate a treatment effect. When little variation in Z appears within levels of X, the data contain little information about the causal effect. The study is not well designed for causal inference.

The assumption of strong ignorability is depicted in Figure 2. We see first a causal arrow between Z and Y as in Figure 1. However, there are also arrows between X and Z and between X and Y, because X contains confounders, that is, predictors of Y that also predict Z. There is also an arrow between U' and Y but not between U' and Z. While U' includes unmeasured predictors of Y these are assumed not related to Z under the assumption of strong ignorability.

Insert Figure 2 About Here

The framework of strong ignorability leads to the following five-step strategy for causal inference when random assignment is impossible:

1. When planning a study, search the literature for theory and prior empirical evidence about the predictors of Y. Measure those as well as possible. Call the measured covariates X, leaving U' unmeasured.
2. Assuming strong ignorability, that is, assuming U' is unrelated to Z given X, compute the mean difference between treatment means within each stratum defined by X; then estimate the average causal effect of Z by averaging these observed mean differences within levels of X.[6]
3. Test the stability of the results by varying which X's are uncontrolled (Rosenbaum, 1999) and

[6] This will typically be a weighted average.

trying different reasonable analytic methods (more below on the analytic options).

4. Test sensitivity (Rosenbaum, 1986). First, hypothesize the unknown association between *U'* and *Z* and examine how inferences about the causal effect change under that hypothesis. Repeat this procedure under a variety of plausible hypotheses.
5. If results are stable over reasonable analytic choices and if the results are relatively insensitive to all plausible associations between *U'* and *Y*, then assign a fairly high level of confidence to the causal inference. Otherwise, the causal inference is more tentative.

To illustrate the logic of sensitivity analysis, suppose, for example, that an exhaustive literature review identified 25 covariates expected by theory or past empirical evidence to predict *Y*. We therefore took pains to measure these 25 *X*s. Now suppose our sensitivity analysis shows that, for our causal inference to be fundamentally wrong, an unmeasured covariate, *U'*, would have to exist that is more confounded with *Z* than any of our 25 *X*s. Then we would likely conclude our results are insensitive to missing confounders, that is, to violations of the strong ignorability assumption. It seems implausible in light of our literature review and painstaking efforts to measure the important *X*s that there remains unmeasured a *U'* that is more "important" than any of our *X*s! This inference of insensitivy assumes reasonably good prior scientific knowledge about the predictors of *Y*.

Step 2 may be implemented in several ways. We consider three: a) linear model-based adjustment; b) stratification by propensity score; and c) inverse-probability-of-treatment weighting.

Linear model-based adjustment. Perhaps the most common strategy for adjusting for *X* is to include the *X*s as covariates in a linear model that also includes *Z* as a predictor. This approach will be satisfactory if a) there are not too many *X*s or the *X*s are not too nearly multicollinear to destabilize the regression estimates; and b) the associations between *X* and *Y* do conform to linearity.

Stratifying by propensity score. An increasingly popular alternative to regression is to stratify the sample on estimated propensity scores. The propensity score is probability of treatment assignment given *X*. When *Z* is binary,

$$propensity\ score = Prob(Z = 1|X). \qquad \textbf{(3)}$$

The propensity score has a remarkable virtue. Suppose a large sample of persons receiving either treatment 1 or 0 are stratified by their propensity scores. Then, within each stratum, the two treatment groups will be balanced in terms of *every covariate* that contributes to the propensity score (see the proof by Rosenbaum and Rubin, 1983a). Thus, under the assumption of strong ignorability, one can construct strata with respect to the propensity score knowing that the mean difference between treatment groups in *Y* within each stratum will be an unbiased estimate of the average causal effect of treatment for persons of the type defined by the values of *X* in that stratum. A weighted average of these stratum-specific treatment effect estimates will then be an unbiased estimate for the average causal effect of the treatment for the population represented by the sample as a whole.

Another advantage of this approach is that an inspection of the distribution of *Z* within each stratum conveys a sense of how much information the data contain about the causal effect of *Z*. When using multiple regression to control *X*, it is easy to overlook the possibility that, within levels of *X*, children tend to be segregated by *Z*, meaning that the data contain little information about the causal effect.[7]

Inverse probability-of-treatment weighting. Stratifying on propensity scores and multiple

[7] Application of a linear model in this setting will create what statisticians call an "extrapolation" (Cochran, 1957). Suppose for example that those in the treatment $Z = 1$ have high values of *X* while those in treatment $Z = 0$ have low values of *X*. Regression can produce the predicted *Y* for a child with $Z = 0$ and a high value of *X* even though no such children exist in the sample. Regression will produce an estimate of the treatment effect for children with high *X*, but that estimate will be based entirely on the extrapolation, which comes from the assumption of linearity in the *X*,*Y* relationship and not form the data.

regression thus constitute two analytic strategies for estimating the treatment effect with adjustment of covariates X. A third approach which has recently gained attention in statistics comes from the theory of sample surveys. Viewing the estimated propensity score for each child as the probability that the child will receive method $Z = 1$, we construct sample design weights (Robins, Hernán, and Brumback, 2000; Hernán, Brumback, and Robins, 2000). The weights we desire are inversely proportional to the pre-assignment probability that the child will receive the treatment that child actually received. Thus, if child i received treatment $Z_i = 1$, we weight that child's outcome, Y_i, inversely proportional to $Pr(Z_i=1|X_i)$. If, on the other hand, child i received treatment $Z_i = 0$, we weight Y_i inversely proportional to the $Pr(Z_i=0|X_i)$. Heuristically, selection bias based on X over-represents certain types of children in each group. We down-weight those so over-represented.

We can alternatively motivate the inverse propensity weights as non-response weights. Within the potential outcomes framework, the counterfactual outcome is a missing datum or "non-response." The inverse-probability-of-treatment weights provide greater weight to those whose X predicts non-response, thus restoring the representativeness of the sample of observed outcomes in reflecting the entire set of outcomes, which include the "non-responding" counterfactual outcomes. Under the assumption of strong ignorability, inverse probability-of-treatment weighting yields a consistent estimate of the treatment effect (see Robins, 2000 for a proof). A consistent estimate is unbiased in large samples. As in the case of propensity-score matching or stratification, one need not assume under this method that a linear regression model holds in using X to predict Y.

Inverse probability-of-treatment weighting becomes an especially important analytic strategy when we turn to the case of studying *sequences* of instructional treatments in the non-randomized case. In that setting, adjusting for X via regression or propensity score matching does not provide satisfactory treatment-effect estimates, even under the assumption of strong ignorability.

In sum, in the case of non-randomized studies of instruction during a single year, there are various ways to conduct the analysis: linear model-based adjustment, propensity score matching or stratification, or inverse probability-of-treatment weighting. The key assumption in all cases is that of strong ignorability conveyed by Equation 2 and Figure 2. We assume that, after properly taking into account *X*, there is no association between unmeasured covariates in *U*' and *Z*. If this assumption holds, then, in principle, there is an analysis that will replicate the results of a randomized experiment. Stability and sensitivity analyses test whether violations of this assumption are plausibly large enough to overturn the key conclusions of the study.

Methods for "non-ignorable" treatment assigment. There are analytic strategies that do not require the assumption of strong ignorability. First, one might find an instrumental variable (c.f., Angrist, Imbens, and Rubin, 1996), a pre-treatment variable that strongly predicts treatment assignment *Z* but that cannot be directly related to *Y*. Logically, the part of *Z* that is predicted by the instrumental varible is unconfounded with all pre-treatment predictors of *Y*. Therefore, the association between this "exogenous" component of *Z* and *Y* reflects the treatment effect. For example, Z can be lottery numbers used for random assignment. If such an instrumental variable exists, the treatment effect can therefore be estimated without bias even if our unmeasured covariates *U*' are related to *Z* given *X*. This can be an effective strategy, but the assumption that the instrumental variable has no direct association with *Y* must be strongly justified theoretically. The data contain no information to evaluate this assumption. Moreover, small departures from the validity of this assumption can produce large biases in the treatment effect estimate (Winship and Morgan, 1999).

Second, one can define a latent (unobservable) variable that determines *Z*. One then constructs a selection model that uses *X* to predict this latent variable. Based on key distributional assumptions, a simultaneous equation model (with the latent variable and *Y* as outcomes) will then provide an unbiased estimate of the treatment effect even though *U*' is associated with *Z* (Heckman, 1979). The difficulty with this approach is that the key distributional assumptions cannot be checked because the latent variable is by definition unobserved. And treatment effect estimates may be sensitive to small departures from the validity of the assumptions.

While methods that do not assume strong ignorability are enticing and sometimes extremely useful (c.f., Little and Yau, 1998), we confine our attention in this paper to methods that assume strong ignorability but that evaluate tentative treatment effect estimates with stability and sensitivity analyses.

Understanding the Causal Effects of Instruction Over More Than One Academic Year in a Randomized Experiment

As mentioned earlier, the effect of a sequence of instructional experiences occurring over multiple years cannot logically be equated to the sum of the effects of instruction occurring each year. Instead, a truly experimental evaluation of the sequence over two years requires that we randomly assign children to one of four sequences, denoted for simplicity $(Z_1,Z_2) = (0,0)$, $(0,1)$, $(1,0)$, or $(1,1)$. A student in such an experiment has a pair of potential outcomes at the end of grade 1 of $(Y^{(0)},Y^{(1)})$ and four more potential outcomes at the end of grade 2. In particular, for a child experiencing $Z_1 = 0$ in grade 1, that child's potential outcomes in grade 2 would be $Y^{(0,0)}$ if assigned to $Z_2 = 0$ in second grade and $Y^{(0,1)}$ if assigned to $Z_2 = 1$ in second grade. In contrast, for a child experiencing $Z_1 = 1$ in first grade, that child's potential outcomes in grade 2 would be $Y_2^{(1,0)}$ if assigned to $Z_2 = 0$ in second grade and $Y_2^{(1,1)}$ if assigned to $Z_2 = 1$ in second grade. Thus each child has six potential outcomes in all. These possibilities are displayed in Figure 3.

Insert Figure 3 About Here

We define the average causal effect of first grade treatment $Z_1= 1$ relative to $Z_1 = 0$ on grade-1 outcomes as

$$E(Y_1^{(1)} - Y_1^{(0)}) = \delta. \qquad \textbf{(4)}$$

We define $E(Y_2^{(1,0)} - Y_2^{(0,0)}) = \Delta_1$ as the causal effect of receiving grade-1 treatment on grade-2 outcome for those not receiving the treatment in grade 2.

Now let us consider the second-grade causal effect. The average causal effect of $Z_2 = 1$ versus $Z_2 = 0$ for those who had experienced $Z_1 = 0$ is

$$E(Y_2^{(0,1)} - Y_2^{(0,0)}) = \Delta_2. \qquad \textbf{(5)}$$

What is the effect of receiving the treatment in both grades 1 and 2? If the treatment effects were simply additive, a child receiving the treatment in both years would benefit by an amount $\Delta_1 + \Delta_2$, the sum of the separate year-1 and year-2 treatment effects on year-2 outcome. If, however, experiencing a second consecutive year of treatment *amplified* (or *muffled*) the effect of the first year treatment, the effects would not be additive. We define the non-additive component as

$$E(Y_2^{(1,1)} - Y_2^{(0,0)}) - (\Delta_1 + \Delta_2) = \Delta^*. \qquad \textbf{(6)}$$

Learning about Δ^* provides a key part of the rationale for studying sequences of instructional experiences rather than limiting our attention to single-year studies.[8] We can express our three causal effects in a single equation,

[8] Suppose the effects of treatment in the two years were additive except for a "fade-out" occurring during the summer. Then we would expect Δ^* to be negative. Twice annual student assessment (fall and spring)

$$E(Y_2^{(Z_1,Z_2)}) = E(Y_2^{(0,0)}) + \Delta_1 Z_1 + \Delta_2 Z_2 + \Delta^* Z_1 Z_2 . \qquad (7)$$

From Equation 7 we can deduce that the effect of grade-1 treatment on year-2 outcome is $\Delta_1 + Z_2\Delta^*$ while the effect of grade-2 treatment on year-2 outcome is $\Delta_2 + Z_1\Delta^*$.

Insert Figure 4 About Here

Figure 4 displays a causal model for an experiment that assigns students at random to a sequence Z_1, Z_2 of instructional treatments. Here U_1 includes "baseline" covariates (covariates measurable before fall of grade 1) that predict Y_1 (and possibly Y_2). These covariates are unrelated to treatment assignment Z_1 and Z_2 which were assigned at random. U_2 are time-varying covariates (e.g., motivation, mental health), that may have been affected by grade-1 treatment Z_1 and that also may be predicted by Y_1 . These time-varying covariates may also predict Y_2. However, these time-varying covariates are not related to Z_2 because assignment to the sequence was randomized prior to the emergence of U_2. Thus, all covariates are independent of treatment assignment, and randomization allows us confidently to assume that treatment assignment is *ignorable*, even in the absence of covariates:

$$U_1, U_2 \perp Z_1, Z_2. \qquad (8)$$

It may, however, be difficult to carry out an experiment in which students are randomly assigned to treatment sequences. A more plausible scenario would involve the random assignment of schools to instructional sequences, where all students within a school receive the same sequence. For example, a school that selects "Success for All," a well-specified instructional intervention, would tend to expose all

would enable us to estimate such an effect directly.

students to a similar instructional intervention over the primary grades.[9] Schools might be assigned at random to receive Success for All versus some other intervention.

However, a study that randomly assigns students to sequences of instruction over two or more years is likely to evolve quickly into a non-randomized study. This will occur primarily because, inevitably, some students will leave a school and others will enter. Such mobility is not under the control of random assignment and may, in fact, be a consequence of the effectiveness of the instruction! It is possible, for example, that parents will decide to remove their child from a school because they view the instructional strategy that school uses as ineffective. Alternatively, they may decide to stay in a school because they admire the instructional approach. Such non-random mobility will undermine the ignorability (Equation 8) that was seemingly conferred by randomization.

Another difficulty with implementing random assignment to sequences of instructional experiences involves the ethics and feasibility of sticking with a pre-determined instructional assignment in grade 2 after one has observed a child's grade 1 progress. Suppose that $Z_1 = 1$ did not seem to work well for a child. This may discourage teachers from sticking with the grade 2 treatment assigned at random. For example, if $Z_2 = 1$ is a logical continuation of $Z_1 = 1$, teachers may wish to avoid giving $Z_2 = 1$ to students who fared poorly under $Z_1 = 1$ while taking pains to insure that children who did well under $Z_1 = 1$ also be assigned to $Z_2 = 1$. In these cases we cannot expect teachers to comply with random assignment.

In other cases, it may be decided that random assignment was simply not possible. Whether a randomized study evolved into a non-randomized study or was never conceived as a randomized study,

[9] Under such an intervention, students would have different instructional experience because programs like Success for All routinely assess student learning and then adapt the instructional strategy to the current level of student proficiency. Thus, students who learn at different rates will encounter different teaching strategies. However, all students experience the same "regime" where an instructional regime is a set of rules for assigning students to instructional strategies given current student status.

Figure 4 and Equation 8 will not hold. Some attempt must be made to identify and control for confounding variables.

Understanding the Causal Effects of Instruction Over More Than One Academic Year in a Non-Randomized Experiment

Suppose now that students are assigned to sequences $Z_{1,}$ Z_2 in grades 1 and 2, respectively, not strictly by chance but instead as a result of a combination factors including parent choice, teacher choice, and a variety of other systematic or accidental factors. Then, as in the case of a single-year study, social scientists take pains to measure a variety of *covariates* that, if ignored, would bias the study. Let X_1 denote "baseline" covariates (covariates measurable prior to fall grade 1), identified as potential confounders; and let X_2 denote "time-varying" covariates (covariates measurable only after fall grade 1) that might be confounded with assignment to the second year's treatment, $Z_2 = 1$ versus $Z_2 = 0$. As in the previous discussion, we assume that, once the observed covariates are taken into account, the unobserved covariates are independent of treatment assignment. In particular, given X_1, unobserved baseline covariates U_1' are independent of treatment assignment Z_1 in grade 1:

$$U_1^{/} \perp Z_1 | X_1. \tag{9}$$

Next, given observed time-varying covariates, X_2, along with other pre-Z_2 observables X_1, Y_1, and Z_1, the unobserved covariates U_1' and U_2' are independent of second-grade treatment, Z_2:

$$U_1^{/}, U_2^{/} \perp Z_2 | X_1, X_2, Y_1, Z_1. \tag{10}$$

In words, the second grade treatment assignment does not depend on unobserved covariates after controlling for the observed covariates, prior outcomes, and treatments.

Assumptions (9) and (10) may be regarded as "sequential strong ignorability" or "sequential randomization" (Robins, 1987), where the randomization at each time point is carried out within levels of prior observed variables.

Insert Figure 5 About Here

These assumptions are displayed graphically in Figure 5. We see that observed pre-treatment covariates X_1 predict both first-grade treatment, Z_1, and outcomes Y_1. However, the unobserved pre-treatment covariates, U_1', while predicting outcomes Y_1, are unconfounded with first-grade treatment, Z_1. The unobserved time-varying covariates, U_2', can predict Y_2 but are unconfounded with second grade treatment Z_2 after taking into account X_1, X_2, Z_1, and Y_1.

The problem of analysis. Our general strategy for the analysis is the same as in the previous section. Assuming sequential strong ignorability (Equations 9, 10), estimate the causal effects. Then assess the stability of the estimates across analytic approaches and the sensitivity of the results to plausible departures from the strong ignorability assumption. As Robins has shown (Robins, Greenland, and Hu, 1999), the problem is that of the three analytic methods mentioned above (linear model-based adjustment, propensity score matching, and inverse probability-of-treatment weighting), only the weighting strategy can be relied upon to produce an unbiased estimate of the effect of sequential treatments. We consider each in turn.

Linear model-based adjustment. To clarify the problem with linear model adjustment, we can modify Figure 5 a) under random assignment to sequences and b) without random assignment but assuming strong ignorability. In each case, we remove U_1' and U_2', the unmeasured covariates that are ignorable under (9) and (10), for simplicity. Under random assignment, we can estimate all causal effects of interest with two simple regression models. First, to estimate δ, the causal effect of $Z_1 = 1$ versus $Z_1 = 0$ on year-1 outcome, we estimate

$$Y_1 = \alpha_0 + \delta Z_1 + e_1. \qquad \textbf{(11)}$$

Now to estimate the effects of grade 2 treatment on grade 2 outcome, we estimate the model

$$Y_2 = \beta_0 + \Delta_1 Z_1 + \Delta_2 Z_2 + \Delta^* Z_1 Z_2 + e_2. \qquad \textbf{(12)}$$

Suppose then that we do not have random assignment. Assuming X_1 is linearly related to Y_1 and that X_1 does not interact with Z_1, we can easily modify Equation 10 by adding X_1 as a covariate. Thus, we have

$$Y_1 = \alpha_0 + \alpha_1 X_1 + \delta Z_1 + e_1. \qquad \textbf{(13)}$$

The coefficient δ is the average causal effect of $Z_1 = 1$ versus $Z_1 = 0$ on Y_1 under these linear model assumptions and under strong ignorability. Thus, estimating the effects of first-grade treatment on first-grade outcomes is straightforward under common assumptions.

But how do we modify Equation 12 to obtain unbiased estimates of the causal effects on second-grade treatment? If we fail to add X_1, Y_1, and X_2 as covariates, we will bias the estimates of the treatment effects because X_1, Y_1, and X_2 are confounded with Z_2. Therefore, we might estimate a model of the form $Y_2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 Y_1 + \alpha_3 X_2 + \Delta_1 Z_1 + \Delta_2 Z_2 + \Delta^* Z_1 Z_2 + e_2$. Unfortunately, if we do control X_2 and Y_1 in the regression, our estimates will be biased by the fact that X_2 and Y_1 are outcomes of the treatment Z_1! Thus, we can use the linear model analysis coherently only if we impose an additional assumption: that X_2 and Y_1 are unconfounded with grade 2 treatment Z_2. In that case, we can include X_1 in Equation 12 as covariates, excluding X_2 and Y_1, yielding the model $Y_2 = \alpha_0 + \alpha_1 X_1 + \Delta_1 Z_1 + \Delta_2 Z_2 + \Delta^* Z_1 Z_2 + e_2$ and obtain an unbiased estimated of the grade 2 treatment effect under linear model assumptions. Note that X_2 and Y_1 have been excluded from the regression based on the assumption that they are uncorrelated with $Z_{1.}$ But this is a strong assumption that can be checked against the data. We find this assumption unsupportable in our illustrative example to be described below. We therefore conclude that the linear model-based adjustment for X_2 and Y_1 provides an unacceptable strategy for implementing the "strong ignorability" analytic strategy. Note we might formulate a pair of simultaneous equations. The first would have Y_1 as an outcome with X_1 and Z_1 as predictors. The second model would have Y_2 as an outcome, with X_1, Z_1, Z_2, Z_1Z_2, $Y_{1,}$ X_2, Y_1Z_2 as predictors. From this pair of simultaneous equations, we might then estimate the total effects of Z_1, Z_2 and Z_1Z_2 from their direct and indirect effects. While somewhat reasonable, this approach would require a host of linear model assumptions, and introduce considerable complexity into the model.

Propensity-score stratification. One might opt to use propensity scores to stratify children as a prelude to estimating within-strata and average treatment effects. But this strategy suffers the same fatal flaw as the linear adjustment strategy described above. Stratifying children according to X_1 will work for estimating the effect of first-grade treatment Z_1 as described in the previous section. However, to estimate the effect of second grade treatment, we face the same dilemma as in the case of linear model adjustment. If we construct propensity scores based on $X_{1,}$ and Z_1, we will not be adjusting for X_2 and Y_I in the estimation of the effect of Z_2 and its interaction with Z_1. On the other hand, if we do include X_2 and Y_I in the estimation of the propensity score, we will bias estimation of the effect of Z_1 and its interaction with Z_2.

Inverse probability-of-treatment weighting. Robins and colleagues (Hernán, Brumback, &

Robins, 2000, 2002; Robins, 2000; Robins, Hernan, and Brumback, 2000) have shown that we can use all prior observables, including $X_{1,}$ Z_1, Y_1, and X_2 in constructing inverse probability of treatment weights that will fully adjust for these observables in estimating the effect of Z_2 and its interaction with Z_1. Robins, Hernán, and Brumback (2000) describe how to compute stabilized weights as recommended. Using this approach, the estimation of the average effects of grade 1 and grade 2 treatments (and their interaction) on grade 2 outcomes follows the simple model of Equations 11 and 12 with the modification that the data are weighted inversely proportional to the estimated pre-assignment probability that a child would receive the treatment actually received.

Illustrative Example: Causal Question, Data, and Methods

The Causal question

For illustration, we consider the relationship between mathematics learning opportunities and learning results in Title I schools. In particular, we explore alternative types of mathematics instructional treatments and their differential effects on the math learning of free-lunch versus non-free-lunch students.

John Carroll's (1963) model framed school learning as a function of instructional processes including opportunity to learn and quality of instruction as well as of student characteristics. Since then, the positive association of instructional time and content coverage with students' academic performance has been well documented by educational researchers (Cooley & Leinhardt, 1980; Fisher, Berliner, Filby, Marliave, Cahen, & Dishaw, 1980; Porter, 1989; Schmidt, McKnight, & Raizen, 1997). In summarizing the research results, Porter and his colleagues (Porter, Floden, Freeman, Schmidt, & Schwille, 1988) concluded that teachers' decisions regarding time allocation to a subject over the course of a school year and content topics to be taught "determine student opportunity to learn, a major influence on student achievement" (p.96).

The above statement strongly implies a causal relationship between teachers' supply of instructional time and content coverage on one hand, and students' learning results on the other hand. However, few, if any, empirical studies in this line of work have undergone rigorous scrutiny from the perspective of causal inference. A typical study of instructional effects is non-randomized and non-longitudinal. Linear model-based adjustment is the most frequently used analytical strategy. Admitting the inadequacy of their data and research methodology, most researchers caution against drawing causal claims from their empirical findings. Despite these major limitations of the previous studies, the positive association between intensive classroom instruction and student achievement is widely presumed to be a causal association and has served as an important theoretical basis for educational policy making in the recent decades.

Scientific knowledge about how instruction affects student learning has been in high demand, especially for Title I and other governmental programs that need to be carried out at the classroom level. Most conventional Title I mathematics programs place major emphasis on drilling arithmetic skills, based on the rationale that disadvantaged students are not ready for learning more advanced mathematics content before they master the basic facts and skills (Secada, 1992). This practice has increasingly received criticism. Many researchers believe that the almost obsessive preoccupation with lower level content is a central problem that explains the poor performance of disadvantaged students (McKnight, Crosswhite, Dossey, Kifer, Swafford, Travers, & Cooney, 1987; Porter, Floden, Freeman, Schmidt, & Schwille, 1988). Others more explicitly assert that these compensatory programs probably widen the poverty gap in mathematics achievement by depriving low-income students of the opportunity to learn any important mathematics (Romberg, 1988).

Challenging instruction for all students is the dominant rhetoric in the current policy discourse (Spillane, 2001). Peterson (1988) found that a mathematics curriculum well designed for typical students showed some outstanding results in classrooms with a large proportion of disadvantaged students. This suggests that poor students may benefit at least as much from well-designed mathematics curriculum and instruction as will their less disadvantaged peers. Based on reasoning from Carroll's conceptual model of school learning, a "well-designed" mathematics program will include features of both adequate instructional time and advanced content that are crucial for low-income students' learning. Adequate instructional time on each single math content topic is necessary for these students, because there are limited resources in their home environment that will assist them in learning the same materials. Without the advantage of being exposed to a variety of mathematics knowledge at home, in the local community, or from summer programs, low-income students also rely on schools for accessing more advanced content topics that will enrich their knowledge structure and prepare them for future study. The causal effect of intensive math instruction versus non-intensive instruction for low-income students' math learning is of particular policy interest in pursuing both equity and excellence.

In this study we estimate the causal effect of a sequence of intensive math instructional treatments on student learning over more than two school years in a non-randomized setting subject to the assumption of sequential strong ignorability of treatment assignment (Equations 9 and 10). One essential feature of instruction is the varying extent to which schools and teachers adjust instruction in response to their perception of students' prior learning experiences and learning needs (Clark & Peterson, 1986; McAninch, 1993; Raudenbush, Rowan, and Cheong, 1993; Shavelson & Stern, 1981). For this reason, we suspect that teachers' selection of instructional treatments can be endogenous to their expectations of the potential student outcomes. As we have elaborated in the previous section, conventional analytical approaches cannot generate a consistent estimate of the causal effect of a later instructional treatment that is endogenous to the results of the previous treatments. Our research tasks include empirically investigating the existence of such endogeneity and applying an alternative approach to handling this particular problem by extending inverse-probability-of-treatment weighting to the multilevel context.

The major goal of this study is to estimate the causal effects of intensive math instruction (i.e., allocating more than average amount of time to math instruction and covering relatively more challenging math content topics) versus less intensive math instruction on students' growth in mathematics knowledge and skills in grades 4 and 5. The estimated results will enable us to address the following research questions: (a) What is the effect of intensive math instruction in grade 4 on grade-4 outcome? (b) What is the effect of this kind of instruction in grade 4 on grade-5 outcome? (c) What is the effect of intensive math instruction in grade 5 on grade-5 outcome? (d) Does the experience of intensive instruction in grade 5 enhance the effect of intensive math instruction received in grade 4? (Equivalently, does the experience of intensive math instruction in grade 4 enhance the effect of intensive math instruction in grade 5?) (e) If intensive math instruction shows a positive effect on average at either or both grade levels, do low-income students benefit as much from the treatment as do other students in this population?

Data

Data for this study were collected by the U.S. Department of Education's Planning and Evaluation Service for the Longitudinal Evaluation of School Change and Performance (LESCP) in 1997, 1998, and 1999 (Westat, 2001). LESCP drew its sample from 67 Title I schools located in 18 school districts in 7 different states. Our sample includes a longitudinal cohort of 4,216 students who progressed from grade 3 to grade 5 during the three study years. These students were assigned to about 190 classrooms in each grade. The students largely came from low-income families with diverse ethnic backgrounds.

We use as a measure of students' math learning the Stanford Achievement Test 9 administered at the end of each of the three study years. Their test scores in different years have been equated on the same scale so that we can assess the learning growth over years. The standard deviation is about 39 points in each year.

All the teachers in each sampled school were asked to respond to a teacher survey questionnaire once a year, which provides information on math instruction in each classroom. Our indicator for intensive instructional treatment is a combination of two measures. One comes from teachers' report on amount of time per week spent on mathematics instruction. Depending on which school and which classroom a student attended, the amount of instructional time on mathematics could be as little as 20 minutes a day or as much as six times that amount. We take a logarithm of this measure and then standardize it within each grade to approximate the diminishing returns of time effects on learning (Brown & Saks, 1987).

A second measure is an index of math content difficulty generated from teachers' annual report on coverage of math content topics ranging from the easiest to the most difficult. Teachers who focused their instruction on more challenging math content topics receive a higher score than those who only covered lower-level topics. The five math content topics that are relatively more difficult are: statistics, algebra (solving 2 equations with 2 unknowns), algebra (solving equation with one unknown), problem solving (distance problems), and computation (operations with fractions). The five less difficult content topics are: measurement (using number lines and rulers), measurement (finding length), perimeter from pictures, numbers and operations (rounding), computation (multi-digit multiplication), and problem solving (word problems using addition, subtraction). The difficulty index is the difference between the number of more difficult topics taught and that of the less difficult topics covered by a teacher within a school year, which is then standardized across all the teachers within each grade level.

We construct a binary treatment variable (Z) for each grade level with $Z = 1$ indicating a teacher’s use of intensive math instruction characterized by emphasis on both instructional time and content difficulty, and $Z = 0$ otherwise. Specifically, Z = 1 if a teacher is above the median on both instructional time and content difficulty within a specific grade level. Note that the cutting-point that we choose for creating this binary measure is grade-specific and cannot be compared across grade levels. The math instruction in a teacher’s classroom is considered intensive in both time and content relative to the instruction in other classrooms at the same grade level. We focus our attention on the instructional treatments that the students received in grade 4 and grade 5. Hence the four possible treatment sequences (Z_1,Z_2) are (0, 0), (0, 1), (1, 0), and (1, 1). Among the 147 grade 4 teachers who taught the student cohort in the middle year and reported their math instructional practices, 36 of them provided intensive math instruction to their students. In grade 5, 58 out of 147 teachers adopted the intensive math instruction in their classrooms.

From student records, teacher survey, and principal survey in each year, we also obtain the following measures.

- Student characteristics: free-lunch status, race/ethnicity, gender, and enrollment in special-service programs such as Title I, LEP, IEP, and migration;
- Teacher characteristics: gender, ethnicity, education level, and teaching experience;
- Classroom composition: class size and proportion of low-achieving students receiving special services;
- School characteristics: school size, proportion of free-lunch children, proportion of minority students, and whether or not the school adopted a school-wide Title I program.

Because of attrition and in-migration, missing data are inevitable in this kind of longitudinal data. A student whose treatment and outcome are both observed in one year is said to have one complete occasion. Among the 4,216 students in the current sample, only 953 of them have complete occasions in all the three study years;1,220 students have two complete occasions; and 2,043 students have only one complete occasion. Table 1 displays the response patterns.

Insert Table 1 about here

We present in Table 2 the descriptive statistics on all of the student, teacher, and school measures. We see that most students in the sample come from low-income families (over 70% of the students are eligible for free lunch). Roughly equal numbers of students are African American and White (44.0% versus 42.9 %), with 10.2% being Hispanic. About two-thirds of the students receive Title I services. At the classroom level, we find class sizes to be modest on average (mean of 14.2). Nearly two thirds of the teachers are White, with a large fraction of African-American teachers (27.4%). Very few teachers are Hispanic. More than 80% of the teachers are female; 42.1% have Masters degrees; and the average number of years of experience is quite high at 13.7. The schools are reasonably typical in size for primary schools (mean enrollment of 433). The averages of the school compositional variables (ethnic composition, free-lunch eligibility, Title I service) quite naturally reflect the overall composition of the student sample.

Insert Table 2 about here

Table 3 shows mean achievement levels by missing data pattern. We note that the average performance of the students who were observed in only one of the three years is consistently lower than that of other students in the sample. The missing data problem for many of these students can likely be attributed to their relatively high mobility. Such students may suffer from discontinuity in math learning as they move from one school to another. It is also possible that some students with missing data were exempted from taking the test in some of the years for reasons such as learning disabilities or lack of English proficiency. This is a strong indication against assuming that the data are missing completely at random. Thus, analysis with a reduced sample excluding students who have one or two-year missing data will be seriously biased (Little and Rubin, 1987). For this reason, we will include all the students in our analysis regardless of their response pattern. We also use missing indicators for estimating our propensity models and construct non-response weights to use in estimating the models for causal effects.

Insert Table 3 about here

Model

Following the logic of causal inference discussed in an earlier section, we propose a theoretical model with potential outcomes for each student in each year. In grade 4, each student has two potential outcomes, associated with intensive math instruction versus non-intensive treatment respectively. Dependent on what treatment a student receives in grade 4 and what the subsequent treatment is in grade 5, the same student has four grade-5 potential outcomes corresponding to the four possible treatment sequences. Our theoretical model makes strong assumptions about the nature of student achievement growth, the structure of treatment effects, and the distribution of random effects defined on students, classrooms, and schools. We assess robustness of all of these assumptions later in a stability analysis that can be regarded as essentially non-parametric.

Our model differs from the simple regression models described in the previous section in two ways. First, rather than using the baseline outcome Y_0 as a covariate, we formulate a person-specific growth model for which Y_0 is the first observation. Treatment effects are then deflections from "expected growth." This model ought to yield treatment effects that are more precisely estimated and more robust to model specification than are treatment effects defined as deflections for an expected post-test. This benefit ought to be especially important for estimating effects of grade 5 treatment. Trajectories from grades 3 to 4 are used to predict counter-factuals for those who experience the grade-5 treatment (see Bryk and Weisberg, 1977; Raudenbush, 2001).

Second, our model reflects the design, according to which time-series data are cross-classified by children and teachers who are, in turn, nested within schools. Recognizing the complexity of the design has two benefits. It makes explicit the classroom-specific nature of the treatment effect and ensures that standard errors will properly reflect the clustered nature of the sample.

Tentatively, assuming a linear growth trajectory for students, we can parameterize the models for the student's potential outcomes as follows for student *i* who encounters teacher *j* in school *k*:

Year 0 (grade 3): $Y_{0ijk} = \beta_{0ik} + v_{0jk} + \varepsilon_{0ijk}$

Year 1 (grade 4): $Y_{1ijk}^{(0)} = \beta_{0ik} + \beta_{1ik} + v_{0jk} + v_{1jk}^{(0)} + \varepsilon_{1ijk}^{(0)}$

$$Y_{1ijk}^{(1)} = \beta_{0ik} + \beta_{1ik} + \delta + v_{0jk} + v_{1jk}^{(1)} + \varepsilon_{1ijk}^{(1)}$$

(14)

Year 2 (grade 5): $Y_{2ijk}^{(0,0)} = \beta_{0ik} + 2\beta_{1ik} + v_{0jk} + v_{1jk}^{(0)} + v_{2jk}^{(0)} + \varepsilon_{2ijk}^{(0,\,0)}$

$$Y_{2ijk}^{(1,0)} = \beta_{0ik} + 2\beta_{1ik} + \Delta_1 + v_{0jk} + v_{1jk}^{(1)} + v_{2jk}^{(0)} + \varepsilon_{2ijk}^{(1,0)}$$

$$Y_{2ijk}^{(0,1)} = \beta_{0ik} + 2\beta_{1ik} + \Delta_2 + v_{0jk} + v_{1jk}^{(0)} + v_{2jk}^{(1)} + \varepsilon_{2ijk}^{(0,1)}$$

$$Y_{2ijk}^{(1,1)} = \beta_{0ik} + 2\beta_{1ik} + \Delta_1 + \Delta_2 + \Delta^* + v_{0jk} + v_{1jk}^{(1)} + v_{2jk}^{(1)} + \varepsilon_{2ijk}^{(1,1)}$$

Here

β_{0ik} is a student-specific intercept defined as the expected math achievement of student *i* in school *k* at the end of year 0, given student-specific random-effects;

β_{1ik} is the student-specific yearly growth rate in math achievement for student *i* in school *k*;

δ is the average year-1 treatment effect on year-1 outcome;

Δ_1 is the average causal effect of year-1 treatment on year-2 outcome (given no treatment in year 2);

Δ_2 is the average causal effect of year-2 treatment on year-2 outcome (given no treatment in year 1);

Δ^*, an interaction effect, is the average increment to year-2 outcome beyond $\Delta_1 + \Delta_2$ that is associated with receiving two consecutive years of treatment;

v_{tjk} is the average increment to a students' math achievement in year *t* that can be attributed to teacher *j*;

ε_{tijk} is the error in predicting student *i*'s math achievement in year *t* given the student-, teacher-, and school-level random effects as well as treatment effects.

For student *i* in school *k*, the student-specific year-1 treatment effect on year-1 outcome is the difference between the two potential outcomes:

$$Y_{1ijk}^{(1)} - Y_{1ijk}^{(0)} = \delta + (v_{1jk}^{(1)} - v_{1jk}^{(0)}) + (\varepsilon_{1ijk}^{(1)} - \varepsilon_{1ijk}^{(0)}), \qquad (15)$$

where $v_{1jk}^{(1)} - v_{1jk}^{(0)}$ is the teacher-specific component of the year-1 treatment effect that is associated with teacher *j* in school *k* who teaches student *i* in year 1; $\varepsilon_{1ijk}^{(1)} - \varepsilon_{1ijk}^{(0)}$ is the student-specific component of the year 1 treatment effect on year-1 outcome. Equation 15 is based on the rationale that, for various reasons, some teachers can be more successful than other teachers in carrying out either treatment or both, and that some students may benefit more than other students from either treatment or both. For simplification, we have assumed that the teacher-specific and student-specific increments to the treatment effect in each year are additive. The student-specific effect of year-1 treatment on year-2 outcome is defined analogously.

Similarly, for student *i* in school *k* who receives no intensive math treatment in year 1, the student-specific year 2 treatment effect on year-2 outcome is:

$$Y_{2ijk}^{(0,1)} - Y_{2ijk}^{(0,0)} = \Delta_2 + (v_{2jk}^{(1)} - v_{2jk}^{(0)}) + (\varepsilon_{2ijk}^{(0,1)} - \varepsilon_{2ijk}^{(0,0)}). \tag{16}$$

The student-specific year-2 treatment effect for student *i* in school *k* who has received intensive math treatment in year 1 is:

$$Y_{2ijk}^{(1,1)} - Y_{2ijk}^{(0,0)} = \Delta_1 + \Delta_2 + \Delta^{*} + (v_{2jk}^{(1)} - v_{2jk}^{(0)}) + (v_{1jk}^{(1)} - v_{1jk}^{(0)}) + (\varepsilon_{2ijk}^{(1,\,1)} - \varepsilon_{2ijk}^{(0,\,0)}). \tag{17}$$

One of the fundamental assumptions for causal inference in standard settings (where outcomes are assumed independent within treatment groups) is the “Stable Unit-Treatment Value Assumption (SUTVA),” which has two components: First, all participants’ outcomes reflect a common set of possible treatment experiences; that is, different persons do not receive different versions of a treatment. Second, the potential outcomes for one individual are independent of the treatment status of other individuals (Rubin, 1978).

Neither of these assumptions is likely to be met in an instructional setting because the treatment is inevitably contingent on the teacher who delivers it; and because the common experience of students in classrooms ensures a mutual effect. In particular, math instruction in elementary schools is mostly organized as a group activity. Typically, teachers decide what treatment to provide to a group of students rather than differentiating the instructional treatment for each individual student. Hence, in most cases, the math learning outcome of student *i* in each year cannot be independent of the treatment that is simultaneously received by other students in the same classroom or even in the same school. In our data set, the nested structure of the data is complex: the repeated measures of math outcomes are cross-classified by students and teachers who are nested within schools. Our theoretical model relaxes the standard SUTVA assumption. It allows teacher-specific treatment effects and models the dependence among responses of students attending the same classroom or school.

Our tentative model conceives of each student as possessing a distinct straight-line learning trajectory given exposure to “average” schools and teachers and given no treatment effects. Teacher effects and school effects are represented as positive or negative deflections from expected student growth curves. A student’s math achievement at the end of each year reflects that student’s personal trajectory plus the cumulative teacher-specific and school-specific deflections to date in addition to the effect of treatment that he or she received (Raudenbush & Bryk, 2002, Chap. 12). The models for the student-specific effects can be written as:

$$\beta_{0ik} = \gamma_0 + u_{0k} + r_{0ik},$$

$$\beta_{1ik} = \gamma_1 + u_{1k} + r_{1ik}.$$

$$\mathbf{u_k} = (u_{0k}, u_{1k})' \sim N(\mathbf{0}, \omega), \tag{18}$$

$$\mathbf{r_k} = (r_{0ik}, r_{1ik})' \sim N(\mathbf{0}, \tau),$$

$$\mathbf{\nu_{jk}} \sim N(\mathbf{0}, \psi^2\mathbf{I}).$$

Here

γ_0 is the population-average math achievement at the end of year 0;

γ_1 is the population-average yearly growth in math achievement;

u_{0k} is school *k*’s effect on math achievement at the end of year 0;

u_{1k} is school *k*’s effect on yearly growth in math achievement;

r_{0ik} is student *i*’s effect on math achievement at the end of year 0;

r_{1ik} is student *i*'s effect on yearly growth in math achievement;

$\mathbf{\nu_{jk}}$ is a vector of teacher-specific deflections.

Again, for the purpose of simplification, we assume that school deflections, teacher deflections, and student-specific increments have constant variance-covariance matrices $\boldsymbol{\omega}$, $\psi^2\mathbf{I}$, and $\boldsymbol{\tau}$, respectively.

Of course we can observe only one of the two potential outcomes in grade 4, and one of the four potential outcomes in grade 5. The unobserved potential outcomes are considered missing. If the students are randomly assigned to treatment sequences, it indicates that we have potential outcomes missing completely at random. Then we can obtain consistent estimates of the average causal effects without adjustment for covariates.

However, since our data are observational rather than experimental, we have reason to suspect that treatment assignment is subject to selection bias at each time point for a variety of reasons. We have come to view the selection process as occurring at the classroom level. Although one might reason that selection occurs at the individual level, our analyses do not support that contention. For example, one might hypothesize that student test scores at the end of grade 3 would predict treatment group membership at grade 4. Our data reveal no association, however, between students' third-grade achievement and their fourth-grade treatment assignment; nor do we find any association between students' fourth-grade achievement and fifth grade treatment assignment. What we do find is that classroom composition in the fall of fourth grade has some association with the treatment assigned to that classroom, with a similar process occurring in fifth grade. We report the results of this classroom-level propensity analysis in the following section of this paper. In sum, the experiment we have in mind is one in which classrooms rather than students are assigned at random to treatments. We view selection bias as also occurring at the classroom level; therefore, our selection model is a classroom-level model.

Given that the instructional treatments under study occur at the classroom level, theoretically, the causal effect of intensive math instruction for classroom *j* in school *k* in year 1, for example, is the difference between the two year-1 potential outcomes of this particular classroom. Each of these two classroom-level potential outcomes is the average of the corresponding student-level potential outcomes.

$$\bar{Y}^{(1)}_{1jk} - \bar{Y}^{(0)}_{1jk} = \delta + (v_{1jk}{}^{(1)} - v_{1jk}{}^{(0)}) + (\bar{\varepsilon}^{(1)}_{1jk} - \bar{\varepsilon}^{(0)}_{1jk}). \tag{19}$$

Our analytic model for the observed outcomes can be stated as

$$Y_{tijk} = \gamma_0 + u_{0k} + r_{0ik} + (\gamma_1 + u_{1k} + r_{1ik})\, t + I_{1t}\delta Z_{1jk}$$

$$+ I_{2t}(\Delta_1 Z_{1jk} + \Delta_2 Z_{2jk} + \Delta^* Z_{1jk} Z_{2jk}) + \sum_{t=1}^{T} \sum_{j=1}^{J_k} D_{tijk} \upsilon_{tjk} + \varepsilon_{tijk}\,. \tag{20}$$

Here I_{1t} is an indicator taking on a value of unity at t=1, 0 otherwise, and I_{2t} is an indicator for time t=2; D_{tijk} is an indicator taking on a value of 1 if student *i* in school *k* has encountered teacher *j* in school *k* by time *t*, 0 otherwise.

For the observed outcomes, the above model is to be weighted by the inverse of a child's estimated probability of receiving the treatment actually received given prior covariates, treatment, and outcomes. Similarly, we create a non-response weight for each student, which is inversely proportional to his or her estimated probability of having the observed response pattern given the observed covariate history. The overall weight is the product of the treatment weight and the non-response weight.

In order to estimate the inverse-probability-of-treatment weight, we compute a propensity score for each classroom in each year, that is, the predicted probability that a classroom will receive intensive math instruction given the measured covariates. The propensity score for grade-4 treatment is a function of the fourth-grade classroom average grade-3 instructional experience, average grade-3 math test score, current class size, proportion of low-achieving students receiving special services, teacher's gender, race, educational level, and teaching experience, in addition to the school characteristics including school size, proportion of students eligible for free-lunch, proportion African American, proportion Hispanic, and whether or not the school adopted school-wide Title I program in year 1. The propensity score for grade 5 treatment is a function of fifth-grade classroom average grade 3 instructional experiences, proportion of students who received intensive math instruction in grade 4, average grade-4 math test score, current math class size, proportion of low-achieving students, as well as teacher characteristics and school characteristics. The results of estimating the second propensity model will also inform us of whether and how the treatment in the later year is endogenous to any measured time-varying factors that can be attributed to the previous year's treatment.

We conduct a further analysis of the effects of intensive math instruction in grade 4 and grade 5 for free-lunch students versus non-free-lunch students. In order to estimate the inverse-probability-of-treatment weight for the sub-populations of interest (free-lunch vs. non-free-lunch in this case), we re-analyze the grade 4 and grade 5 propensity models at student level instead of classroom level, entering free-lunch status as the additional student-level covariate predicting a student's treatment assignment. This allows us to model the marginal distribution of the yearly outcome as a function of time, treatment sequences, as well as students' free-lunch status (Hernán, Brumback, Robins, 2002).

Illustrative Example: Results

In this section, we report the results of our analysis with the LESCP data. We first present the results from estimating the propensity models that provide information about the confounding effects of the baseline and time-varying covariates.

Propensity results

The propensity models are estimated at the classroom level with logistic regression. Table 4 lists for each predictor its regression coefficient estimate, standard error, and *p*-value.

Insert Table 4 about here

If we adopt a nominal significance level of $\alpha = .05$, there are only a few variables that show significant association with treatment assignment. For the grade 4 intensive math instruction, classrooms composed of students who on average have been exposed to relatively more difficult math content topics in grade 3 are more likely to have this treatment in grade 4 on average. However, intensive math instruction is less likely to be adopted in a classroom in which a larger proportion of students' grade 3 math instruction information is missing. Many of these students tend to be migrants who had received math instruction in a different school or district in the previous year. This may indicate that, everything else being equal, teachers tend to be reluctant to use intensive math instruction when student mobility is high. In terms of teacher characteristics, male teachers on average are less likely than female teachers to teach mathematics in an intensive way. However, note that male teachers compose less than 20% of the math teachers in this sample. Classrooms located in schools with a smaller enrollment and with a larger proportion of Hispanic students are more likely to have intensive math instruction. Finally, classrooms that miss information on more than one of the covariates are less likely to adopt the intensive math treatment. These tend to be classrooms whose teachers failed to respond to some of the survey items.

In grade 5, class size is a significant predictor for treatment adoption. Intensive math instruction is less likely to be found in relatively smaller classes in general. We have conducted additional analysis that shows a negative correlation between class size and proportion of low-achieving students receiving special services (r = -.42, p < .001). Class size also shows a significant positive correlation with class average pretest score (r = .29, p < .001). In general, smaller classes tend to have a concentration of low-achieving students. The above information is sufficient to make us wary of the fact that the grade 5 treatments are endogenous to these time-varying covariates that may be outcomes of the grade 4 treatments. Finally, unlike the result of the grade 4 propensity model, grade 5 classrooms that miss information on more than one covariates are more likely to use the intensive math instruction, when we hold everything else constant.

Despite the large number of covariates entered in each of these two propensity models, neither of them shows strong explanatory power. Overall, the grade 5 model is even weaker than the grade 4 model in predicting treatment assignment. The proportion of area under the ROC curve is only .79 for the grade 5 analysis, compared to a c-value of .83 for the grade 4 analysis.

Causal analysis results

We estimate the model specified by Equation 20 via maximum likelihood using the EM algorithm and inverse probability of treatment weighting. For comparison, we also compute the unweighted results. We comment on the causal effects and also on the variance-covariance estimates that are of interest (Table 5).

Causal effects. The second panel of Table 5 presents the results using inverse-probability-of-treatment weighting. Our point estimate for the causal effect of grade-4 treatment on grade-4 outcome is positive, $\hat{\delta}$ = 3.09, but not statistically significant, t = 1.21. Nor do we find a significant effect of grade-4 treatment on grade-5 outcome (for those not receiving grade 5 treatment), $\hat{\Delta}_1$ = 0.04, t = 0.01. We do, however, find a significant effect of grade-5 treatment on grade-5 outcome (for those not receiving grade-4 treatment), $\hat{\Delta}_2$ = 7.52, t = 3.12. There is no evidence that the impact of the grade-5 treatment depends on having had grade-4 treatment, $\hat{\Delta}^*$ = -0.28, t = -0.06. A simple summary of results is therefore possible: there is a significant positive effect of grade-5 treatment, with no evidence of effects of that grade-4 treatment on grade-4 or grade-5 outcome. These results are displayed graphically in Figure 6. The expected gains during grade 4 for all four groups are nearly identical. The grade-5 gains, however, are quite different. Those who receive grade-5 treatment display greater gains during grade 5 than those who do not, regardless of grade-4 treatment assignment.

The unweighted results are very similar to the weighted results (first panel of Table 5). Thus, adjustment for pre-treatment covariates via weighting has little impact on inferences. This result is not too surprising, given that the covariates were found weakly related to treatment assignment. We shall see below, however, that adjustment for covariates has a greater impact on inferences when the grade-4 and grade-5 data are estimated separately, a finding we take up in the discussion section.

Variance-covariance components. We consider how variation in student status is partitioned within and between schools, how variation in rates of growth is partitioned, the contribution of classroom variation, and variation within students. While not central to the causal analysis, these results may have implications for future research.

We estimate that more of the variation in "true" achievement at grade 4 (grade-4 "status") lies between students within schools, $\hat{\tau}$ (π_{00}) = 781.22, than between schools, $\hat{\tau}$ (β_{00}) = 179.25. We conclude that about 179/(179+781) = 19 % of the variance in status lies between schools after controlling treatment effects and adjusting for covariates. The school share of variance is larger, however, when we consider variation in growth rates. The variation in growth rates within schools is estimated to be $\hat{\tau}$ (π_{11}) = 51.60 versus variation between schools $\hat{\tau}$ (β_{11}) = 30.66, so that the between-school share is about 31/(31+52) = 37%. This share looks much larger in the unweighted results.

A large component of variation lies between classrooms, $\hat{\psi}^2 = 168.97$. Recall that classrooms are time varying. It may be that these time-varying "deflections" in student achievement reflect, at least in part, aspects of instructional practice not represented in our characterization of instructional treatments. The within-child variance of $\hat{\sigma}^2 = 254.02$ represents measurement error in the outcome plus deviations from linearity not explained by the treatment effects and classroom effects. We take up the issue of non-linearity in our stability analyses.

Insert Table 5 about here

Stability Analysis

We assess stability of results using an analysis that relaxes key assumptions of the theoretical model (Equations 14 and 20). First, we estimate separate average quadratic trajectories for each treatment sequence. This allows non-linear differences between treatment sequences not attributable to treatment. It also may reveal undetected selection bias to differentiate the intercepts of the four treatment groups. The results are displayed in Figure 7. The two groups receiving grade-5 treatment display near-straight-line growth. The two groups not receiving treatment show evidence of a negative curvature. This difference creates an expected gap that is very close to the treatment effect estimate of 7.52 estimated earlier (second part of Table 5).

We also use data from grades 4 and 5 to compute estimates of the grade-5 treatment effect under four different models (Table 6). A naïve model failing to adjust for any covariate produced a point estimate of 12.70. A model using the propensity score as a covariate reduced this estimate to 7.17, close to our original estimate of 7.52. A model using propensity-score stratification produced an estimate of 8.11 (pooled within the five strata). The discrepant estimate of 3.40 is based on a model that used the grade-4 outcome as an additional covariate. Based on the results of the more complete analyses (e.g., Figure 7), we are inclined to view this covariance adjustment as flawed based on a comparison to models that fully exploit grade-3 outcome data.

Insert Table 6 about here

Sensitivity Analysis

In this part of the analysis we test the sensitivity of inferences to plausible departures from the assumption of strong ignorability. To study the sensitivity of our inference about grade-5 treatment effect, we imagine an unobserved pre-treatment covariate, *U*, having the following characteristics: (a) The standardized mean difference in *U* between treatment groups $Z_2 = 1$ and $Z_2 = 0$ is as large as the largest standardized mean difference observed for any covariates *X* in Table 4; and (b) the linear association between *U* and Y_2 is as large as the largest linear association of any covariate in Table 4. We find that the grade-5 treatment groups differ in class size by .48 standard deviation units, the largest difference observed. Among all the covariates, proportion of African American students in school has the largest association with Y_2, its standardized regression coefficient being .152. If *U* had as strong associations with Z_2 and Y_2 as these two covariates, our confidence interval would be (0.45, 9.67) rather than the obtained interval of (2.80, 12.24). We conclude that our results are not highly sensitive to omission of *U*. We do note, however, that as the association between *U* and Y_2 grows beyond .15, we find confidence intervals that contain zero.

Insert Table 7 about here

Differential Treatment Effects

Finally, we examine if the treatment effects vary between free-lunch and non-free-lunch students. We find that a student not eligible for free-lunch who receives the intensive math treatment in grade 4 is expected to gain 6.49 more points in the grade 4 math test. This value is close to being significantly different from zero with a *t* ratio of 1.93. Its 95% confidence interval is (-0.11, 13.09), heavily leaning toward the positive side. For a non-free-lunch student who receives no intensive math treatment in grade 5, the grade 4 treatment shows an effect of as much as 11.28 points on the grade 5 outcome. However, the standard error of this estimate is especially large (6.16) due to the small number of students having this particular treatment sequence. Consequently, its confidence interval ranges from –0.79 to 23.35. The average benefit of grade 5 intensive math treatment alone for a non-free-lunch student who receives no intensive math treatment in grade 4 is estimated to be 9.47 with a confidence interval of (2.75, 16.18). There is no evidence that the effect of grade-5 treatment depends on the grade-4 treatment for these students.

The results for the free-lunch students are slightly different. The effect of grade-4 intensive math treatment is 0.38 points on the grade 4 outcome, and is –4.89 points on the grade 5 outcome. Neither estimate is significantly different from zero. Nonetheless, the grade-5 intensive math treatment shows a positive effect of 7.96 points on the grade 5 outcome for free-lunch students who receives no intensive math treatment in grade 4. Again, we find no evidence that the effect of grade-5 treatment depends on the grade-4 treatment for these students.

Insert Table 8 about here

Overall, we find some tentative evidence that the grade-4 intensive math treatment benefits the non-free-lunch students, while the grade-5 intensive math treatment shows significantly positive contributions to the math learning of both free-lunch and non-free-lunch students in the Title I schools. See Figure 7 for an illustration of the growth trajectories for free-lunch versus non-free-lunch students in each of the four treatment sequences.

Insert Figure 7 about here

Some exploratory analyses with this data set shows that free-lunch students have a significantly higher chance of receiving the intensive math treatment compared with non-free-lunch students in grade 5 ($X^2_{(2)} = 16.50$, $p < .001$), while the probability of receiving the intensive math treatment in grade 4 is the same for both groups ($X^2_{(2)} = 1.96$, $p = .376$). We further investigate the allocation of math instructional time and content coverage by both free-lunch status and grade level. We find no differentiation in grade-4 math instruction between free-lunch students and non-free-lunch students in either the treatment group or the control group. However, in grade 5, free-lunch students receive substantially more instructional time and are exposed to more advanced math content topics than their non-free-lunch counterparts are in the treatment group (for instructional time, $t = 3.24$, $p < .001$; for content coverage, $t = 5.66$, $p < .001$), while such differentiated instruction is not found in the control group. In light of the estimation results that free-lunch students benefit from the grade 5 treatment, this finding may suggest that improving disadvantaged students' math learning requires more investment of instructional resources in terms of both time and content difficulty than what is typically needed by their non-free-lunch peers. This is a tentative hypothesis to be tested in our future research.

Conclusions

Motivating this work is our belief that studying the causal effects of alternative instructional approaches lies at the heart of any attempt to increase the scientific quality and utility of educational research. We have argued that the utility of research on school governance and school resources depends on the health of an ongoing program of research on the causal effects of classroom practice. However, this is a challenging problem not only because of the difficulties of conducting rigorous field research in classrooms during a single year. Our understanding of the impact of instruction depends ultimately on making inferences about the effects of sequences of instructional effects. Experiments that assign students to alternative multi-year sequences of instruction are difficult to sustain; and inference based on non-experimental studies must overcome problems of endogeneity of treatment assignment not encountered in single-year studies. We offer the following conclusions about modeling the causal effects of multi-year sequences of instruction from non-experimental data.

1. A promising approach is to collect data on relevant pre-treatment and time-varying covariates. Tentative assessment of causal effects is based on the assumption of strong ignorability: that treatment assignment at any given time is unrelated to unobserved pre-treatment data. One then assesses the sensitivity of these estimated treatment effects to plausible departures from the strong ignorability assumption.

2. However, standard methods of analysis, including linear model adjustment and propensity score stratification, do not provide consistent estimates of time-varying treatment effects even under the strong ignorability assumption. We have demonstrated how a weighting approach developed by Robins and his colleagues in epidemiology can yield adjustment for pre-

treatment observables that will produce consistent estimates of average treatment effects. This approach uses inverse-probability-of treatment weights to adjust for measured covariates and yields consistent estimates of average treatment effects under the strong ignorability assumption as well as the assumption of nonzero probability of receiving alternative treatments for each observed covariate pattern.

3. The above method requires adaptation, however, in the context of instructional research. Instructional treatments are delivered by teachers and mediated by interactions occurring among students. This means that students will experience a different version of the treatment depending on the teacher to whom they are assigned and also depending on their classmates. The strategy that we have proposed represents classroom-specific contributions to the causal effect as random time-varying effects in the model. Interpretation of the associated variance components is also conditional on strong ignorability.

We illustrated the proposed approach using longitudinal data to estimate the effects of intensive mathematics instruction in fourth and fifth grades. Our findings include:

1. The intensive math instruction had significant positive effects in grade 5.

2. Effects of grade-4 treatment on grade-4 and grade-5 outcomes were negligible.

3. The results based on inverse-probability-of-treatment weighting were quite similar to the unweighted results that failed to adjust for most of the pre-treatment confounders. This similarity is consistent with another key finding, that is, models using pre-treatment observables to predict treatment group membership have generally weak explanatory power.

4. Inferences about grade-5 treatment effects were moderately insensitive to plausible confounding effects of unmeasured covariates.

5. Other intriguing results include a finding of substantial variation between schools in average growth in mathematics and also large differences between classrooms in annual increments to mathematics achievement. These findings suggest that further investigation of the causes of school and classroom differences in effectiveness might well be promising.

6. There was some tentative evidence that non-poor children benefit from intensive math instruction in grade 4. Both the poor and non-poor children appear to benefit from intensive math instruction in grade 5.

Our final comments concern lessons we have learned about design and measurement in studies of instructional effects. First, a major strength of the LESCP data is its longitudinal character. The repeated measures enable us to conceive of treatment effects at a given grade as deflections from a child-specific growth trajectory estimable from the repeated measures data. For example, we were able to estimate grade 5 treatment effects with good precision and, presumably, a minimum of bias because the data allowed adjustment for pre-treatment *growth* rates as well as pre-treatment status. Multiple pre-treatment measures provide a strong basis for quasi-experimental design (Bryk and Weisberg, 1977; Raudenbush, 2001). The availability of a longitudinally equated metric for student achievement was essential in our growth models. Perhaps our chief concern in the illustrative example was the measurement of the treatment, taken by teacher self-reports of instructional practice. We transformed these data to construct a binary indicator of teaching practice consistent with current thinking about mathematics instruction. A critic of our work may justifiably speculate, however, that measurement error on the causal variable attenuates the estimate of the causal effect of treatment. We would have strongly preferred to study a deliberately planned intervention verified by observational means to have been implemented with some fidelity.

We wish to thank the department of Education's Planning and Evaluation Service for sharing these data and providing support for the reanalysis. Willingness to share program evaluation data with the scientific community is an essential step in improving the scientific quality of educational research.

References

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-472.

Boruch, R., & Mosteller, F. (2002). *Evidence Matters: Randomized trials in education research.* Brookings.

Brown, B. W., & Saks, D. H. (1987). The Microeconomics of the Allocation of Teachers' Time and Student Learning. *Economics of Education Review*, *6*(4), 319-32.

Bryk, A. & Weisberg, H. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin*, *84*(5), 950-962.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In Merlin C. Wittrock (Ed.), *Handbook of Research on Teaching: A Project of the American Educational Research Association* (3rd ed.) (pp.255-296). New York: Macmillan Publishing Company.

Cochran, W. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*(3), 261-281.

Cohen, D., K., Raudenbush, S. W., & Ball, D. L. (2002). Resources, instruction and research. In R. Boruch & F. Mosteller (Eds.), *Evidence Matters: Randomized trials in education research.* Brookings.

Cook, T., Habib, F., Phillips, M., Settersten, R., Shagle, S., & Degirmencioglu, S. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543-597.

Cook, T., Murphy, R., & Hunt, H. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal, 37*(2), 535-597.

Cooley, W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational evaluation and policy analysis, 2*, 7-25.

Englert, C. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal, 28*(2), 337-372.

Fisher, C., D. Berliner, N. Filby, R. Marliave, L. Cahen, & M. Dishaw. (1980). Teaching behaviors, academic learning time, and student achievement: An overview. In C. Denham and A. Lieberman (Eds.), *Time to learn*. Washington, DC: National Institute of Education.

Finn, J. D., & Achilles, C. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577.

Finn, J. D., & Achilles, C. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97-109.

Grouws, D. (1992). *Handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics.* New York: MacMillan.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161.

Hernán, M., Brumback, B., & Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology, 11*(5), 561-570.

Hernán, M., Brumback, B., & Robins, J. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Staistics in Medicine, 21*, 1689-1709.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons.

Little, R., & Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*(2), 147-159.

McAninch, A. R. (1993). *Teacher thinking and the case method: Theory and future directions*. New York: Teachers College Press.

McKnight, C.C., Crosswhite, J. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective*. Champaign, IL: Stipes.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 307-354). Washington DC: American Educational Research Association.

Peterson, P. L. (1988). Teachers' and students' cognitional knowledge for classroom teaching and learning. *Educational Researcher*, *17*(5), 5-14.

Peterson, P., & Hassel, B. (1998). *Learning from school choice.* Washington DC: The Brookings Institution.

Pinnell, G., DeFord, D., Lyons, C., & Bryk, A. (1995). Comparing instructional models for the literacy education of high-risk first graders: Reply. *Reading Research Quarterly, 30*(2), 272-275.

Porter, A. (1989). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher*, *18*(5), 9-15.

Porter, A., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (Vol. 1, pp.96-113). Hillsdale, NJ: Erlbaum.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data, *Annual Review of Psychology*, *52*, 501-525.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks: Sage Publications.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal, 30*(3), 523-553.

Richardson, V. (2001). *Handbook of Research on Teaching (4th Edition).* Washington DC: American Educational Research Association.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases, 40*(2), 139S-161S.

Robins, J. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. Elizabeth Halloran and Donald Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp.95-134). New York: Springer.

Robins, J., Hernán, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550-560.

Robins, J., Greenland, S., & Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of repeated binary outcome. *Journal of the American Statistical Association, 94*(447), 687-700.

Romberg, T. A. (1988). Mathematics for compensatory school programs. In B. I. Williams, P. A. Richman, & B. J. Mason (Eds.), *Designs for Compensatory Education: Conference Proceedings and Papers*. Washington, D. C.: Research and Evaluation Associates.

Rosenbaum, P. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*(385), 41-48.

Rosenbaum, P. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*(3), 207-224.

Rosenbaum, P. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, *14*(3), 259-304.

Rosenbaum, P., & Rubin, D. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika, 17,* 41-55.

Rosenbaum, P., & Rubin, D. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B, 45,* 212-218.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66,* 688-701.

Rubin, D. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1-26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6,* 34-58.

Schmidt, W. H., McKnight, C. C. & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht: Kluwer Academic Publishers.

U.S. Department of Education. (February, 2002). *Scientifically Based Research Seminar*. Sponsored by Assistant Secretary Susan B. Neuman and the Office of Elementary and Secondary Education, U.S. Department of Education.

Secada, W. G. (1992). Race, ethnicity, social class, language, and achievement in mathematics. In Douglas A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning*. New York: MacMillan Publishing Company.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51,* 455-498.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children* (C. E. Snow, M. S. Burns, & P. Griffin, Eds.). Washington D.C.: National Academy Press.

Spillane, J. P. (2001). Challenging instruction for "all students": policy, practitioners, and practice. In Susan H. Fuhrman (Ed.), *From the Capitol to the Classroom: Standards-Based Reform in the States*. Chicago, IL: National Society for the Study of Education: Distributed by the University of Chicago Press.

Westat (2001). The Longitudinal Evaluation of School Change and Performance (LESCP) in Title I Schools. Washington, DC: Us Department of Education, Planning and Evaluation Services, DOC #2001-20.

Winship, C., & Morgan, S. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659-707.

Wittrock, M. (1986). *Handbook of research on teaching (3rd edition).* New York: MacMillan.

Table 1. Sample Response Pattern

Observations	**Grade 3**	**Grade 4**	**Grade 5**	**Total**
a. All the three years	953	953	953	2859
b. Grade 3 and 4 only	730	730		1460
c. Grade 3 and 5 only	127		127	254
d. Grade 4 and 5 only		363	363	726
e. Grade 3 only	1490			1490
f. Grade 4 only		435		435
g. Grade 5 only			118	118
Total	3300	2481	1561	7342

Table 2. Descriptive statistics

Variable	Mean	SD
Students (I = 4,216)		
Free lunch (1 = Yes; 0 = No)	.723	.447
African American (1 = Yes; 0 = No)	.440	.496
Hispanic (1 = Yes; 0 = No)	.102	.303
White (1 = Yes; 0 = No)	.429	.495
Other ethnic groups (1 = Yes; 0 = No)	.029	.168
Gender (1= Male; 0 = Female)	.492	.500
Title I (1 = Yes; 0 = No)	.672	.470
Grade 3 math content difficulty	-.138	.951
Grade 3 math instructional time	.059	.910
Teachers/Classrooms (J = 386)		
Grade (1 = Grade 4; 0 = Grade 5)	.497	.501
Gender (1 = Male; 0 = Female)	.191	.394
African American (1 = Yes; 0 = No)	.274	.447
Hispanic (1 = Yes; 0 = No)	.017	.129
White (1 = Yes; 0 = No)	.659	.475
Other ethnic groups (1 = Yes; 0 = No)	.051	.220
Degree (1 = Master's or above; 0 = Bachelor's or below)	.421	.494
Teaching experience	13.670	9.727
Class size	14.165	6.554
Proportion of low achievers	.261	.338
Intensive math treatment (1 = Yes; 0 = No)	.320	.467
Schools (K = 67)		
Year 1 school size	432.299	146.067
Year 1 % free-lunch	.731	.191
Year 1 % African American	.405	.390
Year 1 % Hispanic	.088	.159
Year 1 school-wide Title I (1 = Yes; 0 = No)	.806	.398
Year 2 school size	440.239	142.641
Year 2 % free-lunch	.726	.188
Year 2 % African American	.421	.392
Year 2 % Hispanic	.092	.159
Year 2 school-wide Title I (1 = Yes; 0 = No)	.821	.386

Table 3. Average Math Achievement and Proportion of Free-lunch by Grade and Response Pattern

Response Pattern	***n***	**Average math achievement**			**% Free-lunch**
		Grade 3	Grade 4	Grade 5	
All three years	953	597.70	621.17	642.26	.67
Grade 3 and 4	730	593.46	616.51		.74
Grade 3 and 5	127	595.50		636.88	.69
Grade 4 and 5	363		611.80	635.96	.63
Grade 3 only	1490	585.28			.76
Grade 4 only	435		605.05		.80
Grade 5 only	118			629.10	.72
Total	4216	591.07	615.38	639.29	.72

Table 4. Propensity Model Results

Predictor	Grade 4 treatment			Grade 5 treatment		
	β	SE (*β*)	*p*	*β*	SE (*β*)	*p*
Average grade 3 content difficulty	1.215	.501	.015	-.654	.421	.120
Average grade 3 math time	-.249	.520	.632	.558	.466	.231
% having grade 4 intensive math	----	----	----	1.171	.656	.075
Average math pretest score	.032	.223	.885	.090	.238	.706
Class size	-.032	.049	.517	.137	.048	.004
% low achievers receiving services	-.379	.993	.702	-1.440	.968	.137
Teacher’s educational degree	.046	.619	.940	.118	.495	.811
Teaching experience	-.046	.031	.144	.007	.024	.761
Teacher’s gender	-1.566	.780	.045	.355	.580	.541
African American teacher	.941	.710	.185	-.092	.588	.876
Teacher of other non-white ethnicity	.761	1.123	.498	.356	.969	.713
School size	-.006	.003	.015	-.001	.002	.701
% free-lunch students in school	-3.392	2.100	.106	.029	.021	.168
% black students in school	.721	1.094	.510	-.569	.994	.567
% Hispanic students in school	4.607	1.989	.021	-2.166	1.890	.252
School-wide Title I program	.043	.869	.960	.817	.896	.362
% missing grade 3 instruction info	-2.969	1.211	.014	-.695	.859	.419
% missing grade 4 treatment info	----	----	----	-1.322	.739	.073
Missing at least one other covariate	-6.799	1.947	.000	3.528	1.792	.049

Table 5. Treatment Effect Estimation Results

	Unweighted Model			Weighted Model		
Fixed effects	**Coefficient**	**SE**	***t***	**Coefficient**	**SE**	***t***
Intercept, γ_0	609.828	1.978	308.361	610.178	2.004	304.456
Growth rate, γ_1	20.934	1.176	17.807	21.253	1.175	18.094
Grade 4 treatment on grade 4 outcome, δ	2.699	2.495	1.081	3.089	2.255	1.209
Grade 4 treatment on grade 5 outcome, Δ_1	0.390	3.661	0.106	0.042	3.834	0.011
Grade 5 treatment on grade 5 outcome, Δ_2	7.819	2.461	3.190	7.518	2.408	3.123
Two-way interaction of grade 4, grade 5 treatment on grade 5 outcome, Δ^*				-0.280	4.577	-0.061
Variance components		**Estimate**			**Estimate**	
Within students						
σ^2		308.841			254.017	
Between students						
τ (π_0)		770.106			781.219	
τ (π_1)		14.999			51.597	
Corr (π_0, π_1)		-.196			-0.134	
Between schools						
ω (β_{00})		171.977			179.245	
ω (β_{10})		30.364			30.660	
Corr (β_{00}, β_{10})		.392			.419	
Between classrooms						
ψ^2 (ν)		171.939			168.970	

Table 6. Stability Analysis Results

		Treatment Effect	Standard Error	*t* Ratio
Model 1	Naïve	12.70	4.15	3.06
Model 2	Propensity as covariate	7.17	3.38	2.12
Model 3	Propensity stratification	8.11	3.67	2.21
Model 4	Propensity stratification & pretest as covariates	3.40	2.66	1.28

Table 7. Sensitivity Analysis Results

Hypothetical standardized mean difference in a covariate	Hypothetical standardized regression coefficient	Adjusted treatment effect	95% CI with standard error equal to 1.972	Minimum standard error for rejecting the null hypothesis
.10	.15	6.61	(2.00, 11.22)	3.37
.20	.15	6.62	(1.61, 10.83)	3.18
.30	.15	5.83	(1.22, 10.44)	2.98
.40	.15	5.45	(0.84, 10.06)	2.78
.50	.15	5.06	(0.45, 9.67)	2.58
.60	.15	4.67	(0.06, 9.28)	2.38
.70	.15	4.28	(-0.33, 8.89)	2.18

Hypothetical standardized mean difference in a covariate	Hypothetical standardized regression coefficient	Adjusted treatment effect	95% CI with standard error equal to 1.972	Minimum standard error for rejecting the null hypothesis
.48	.10	5.77	(1.16, 10.38)	2.95
.48	.20	4.55	(-0.06, 9.16)	2.32
.48	.30	3.32	(-1.29, 7.93)	1.69
.48	.40	2.09	(-2.52, 6.70)	1.07
.48	.50	0.87	(-3.74, 5.48)	0.44

Table 8. Weighted Estimates of Treatment Effects As a Function of Child Poverty

	Non-Free-Lunch Students			Free-Lunch Students		
Fixed effects	**Coefficient**	**SE**	***t***	**Coefficient**	**SE**	***t***
Intercept, γ_0	618.506	2.257	274.083	607.523	2.040	297.788
Growth rate, γ_1	20.138	1.298	15.515	21.435	1.252	17.123
Grade 4 treatment on grade 4 outcome, δ	6.487	3.368	1.926	0.378	2.765	0.137
Grade 4 treatment on grade 5 outcome, Δ_1	11.283	6.164	1.830	-4.893	4.465	-1.096
Grade 5 treatment on grade 5 outcome, Δ_2	9.466	3.425	2.764	7.957	2.570	3.096
Two-way interaction of grade-4 and grade-5 treatments on grade-5 outcome, Δ^*	-4.683	7.582	-0.618	3.928	5.230	0.751

Variance components	**Estimate**
Within students	
σ^2	258.044
Between students	
τ (π_0)	762.354
τ (π_1)	49.109
Corr (π_0, π_1)	-.134
Between schools	
ω (β_{00})	162.284
ω (β_{10})	33.909
Corr (β_{00}, β_{10})	.424
Between classrooms	
ψ^2 (ν)	163.902

Figure 1. Causal Effect of *Z* on *Y* in a Single-year Randomized Study

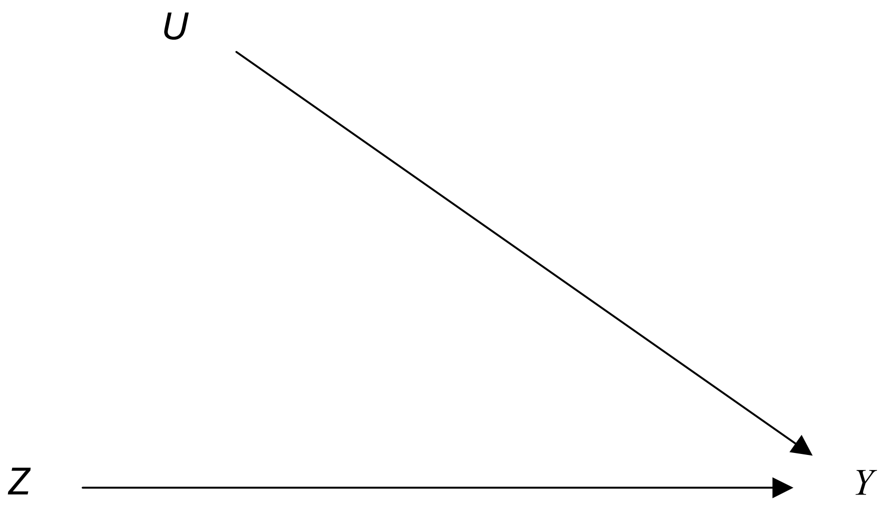


Figure 2. Causal Effect of *Z* on *Y* in a Single-year, Non-Randomized Study with Strongly Ignorable Treatment Assignment

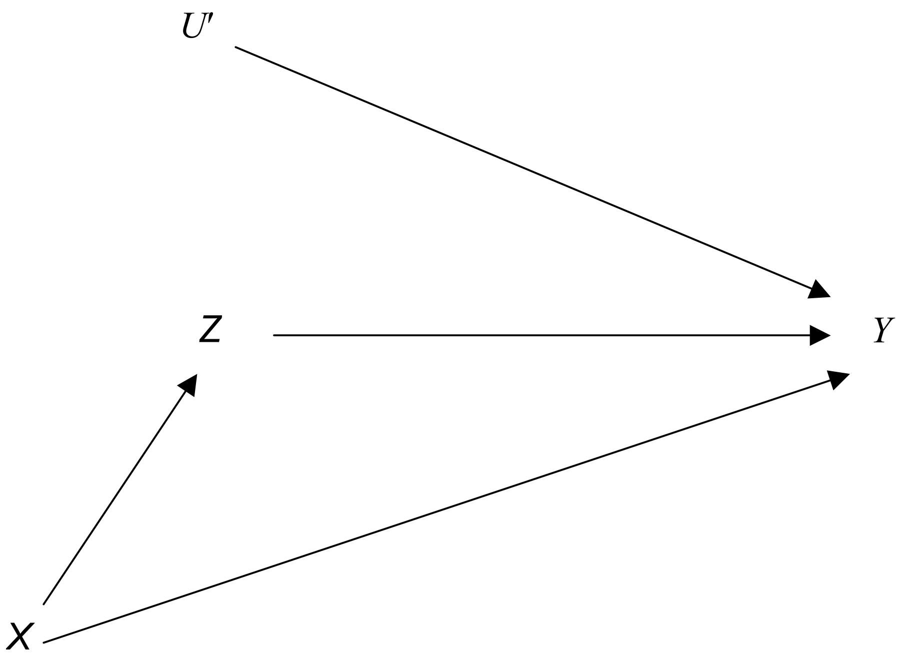


Figure 3. Potential Outcomes in a 2-year Study of Binary Treatments, *Z_1* and *Z_2*

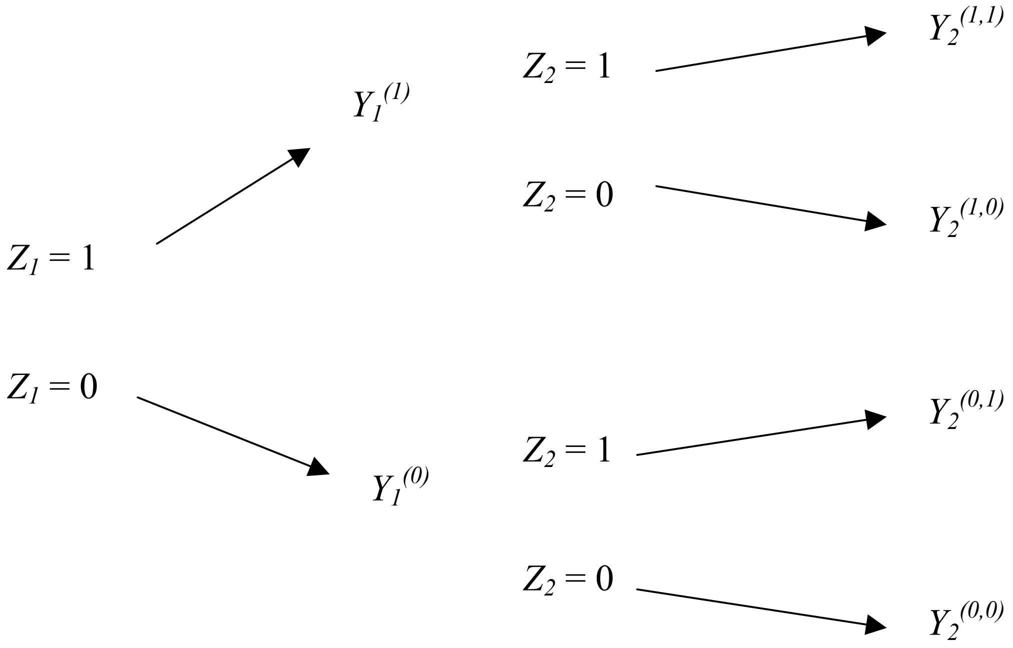


Figure 4. Causal Effects of Z_1, Z_2 in a Randomized 2-year Study

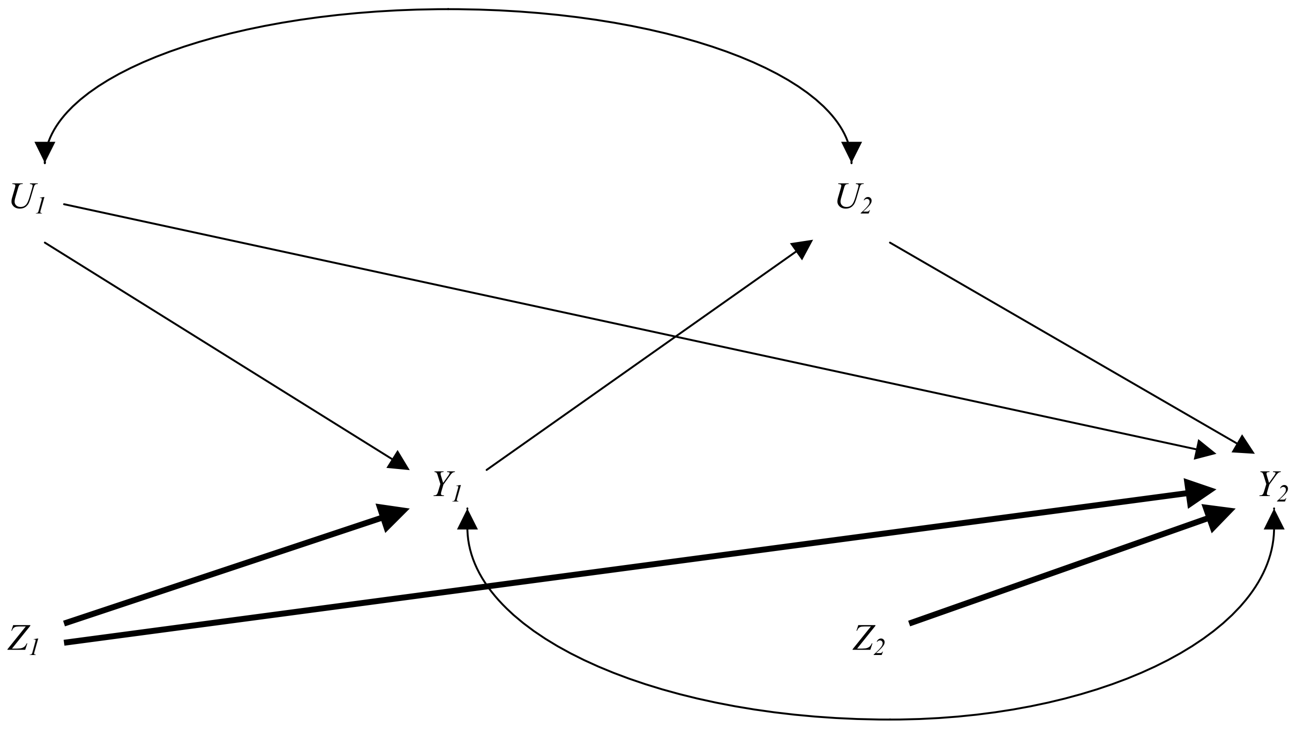


Figure 5. Causal Effects of Z_1, Z_2 in a Non-Randomized 2-year Study Assuming Strongly Ignorable Treatment Assignment

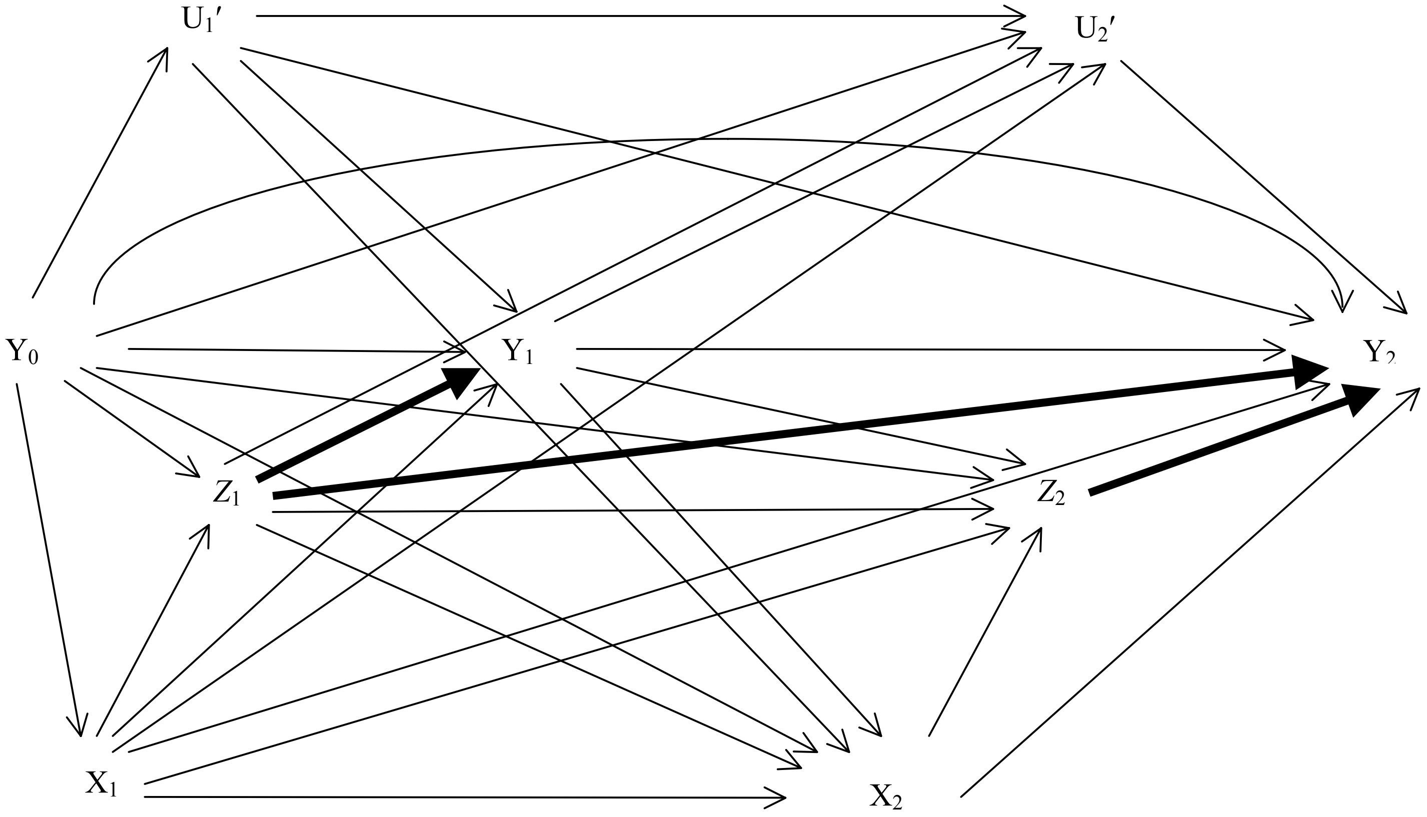


Figure 6. Predicted Treatment Effects on Linear Growth Trajectory

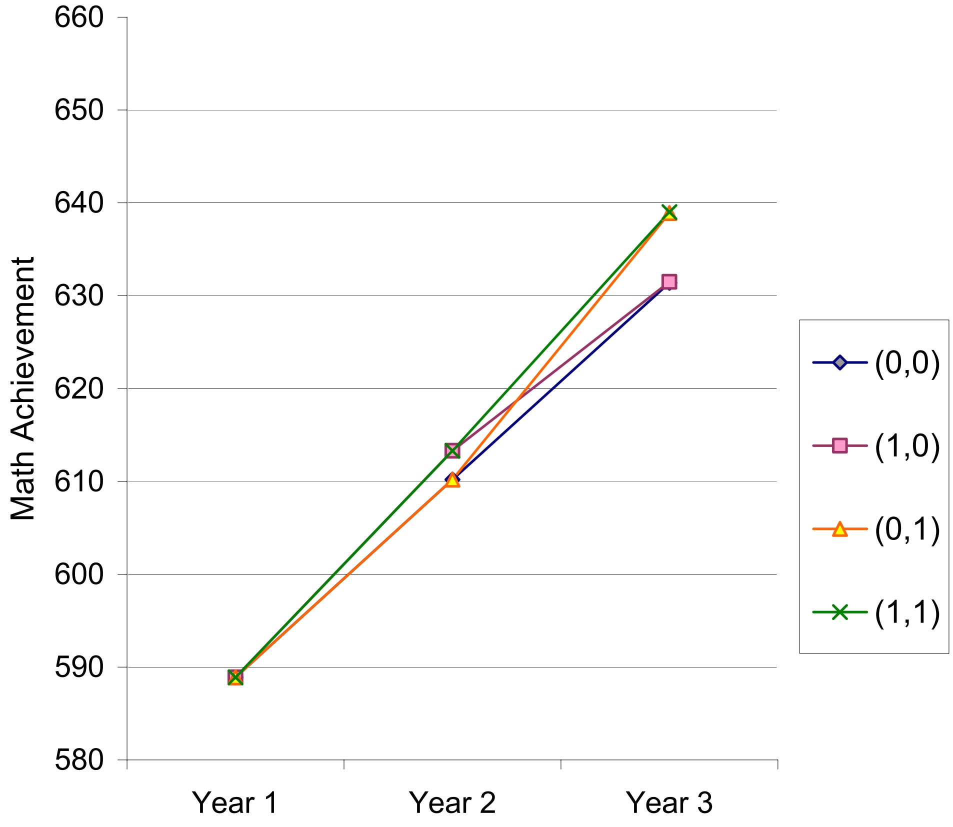


Figure 7. Quadratic Growth Trajectory for Each Treatment Sequence

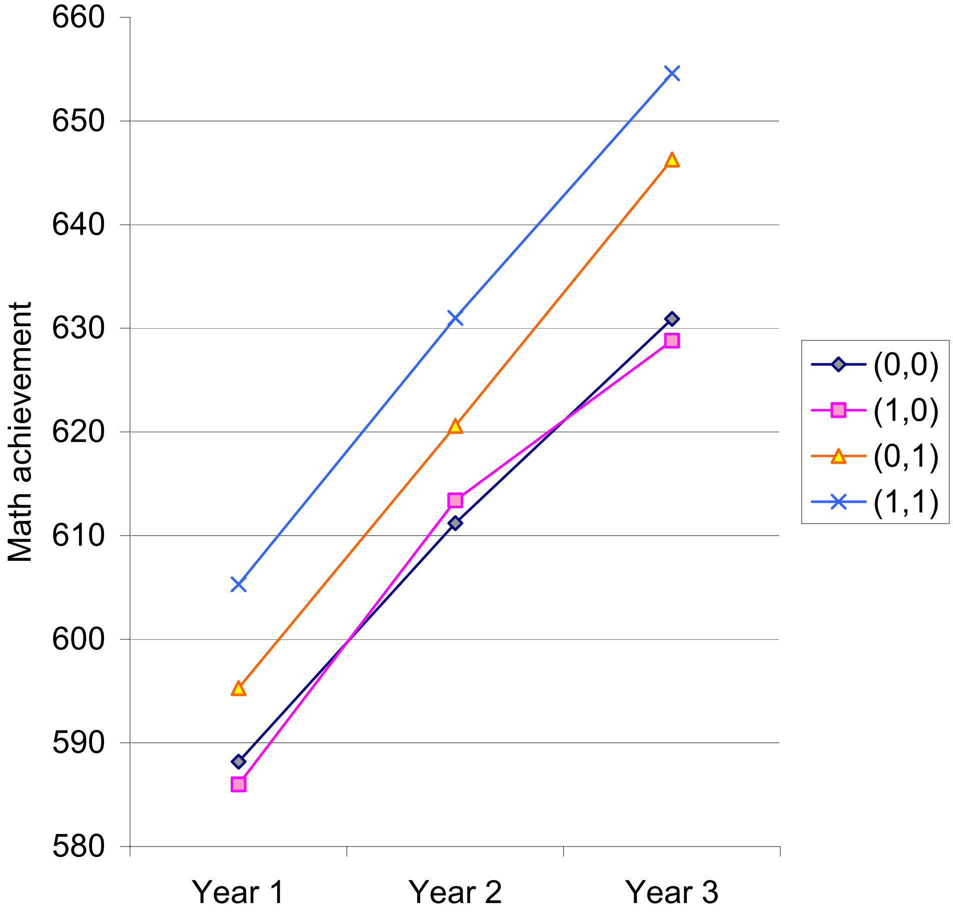
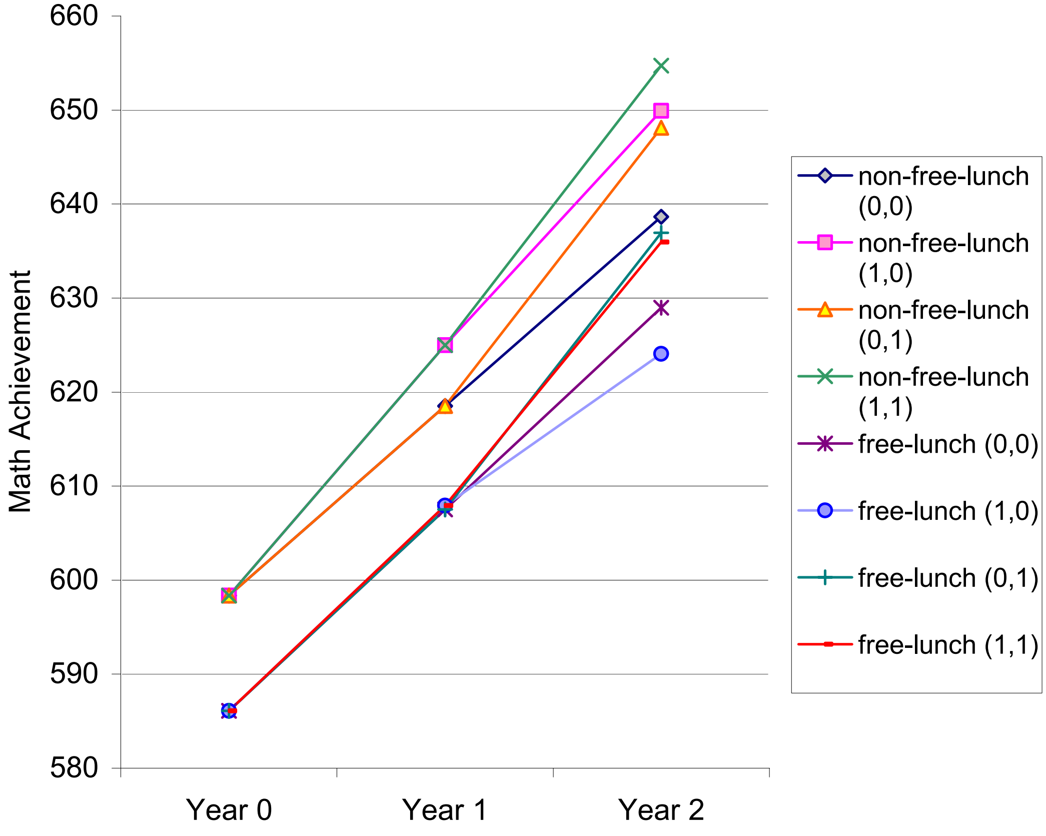


Figure 8. Predicted Treatment Effects on Linear Growth Trajectory for Free-Lunch vs. Non-Free-Lunch Students


Math Achievement
Year 0
Year 1
Year 2
non-free-lunch (0,0)
non-free-lunch (1,0)
non-free-lunch (0,1)
non-free-lunch (1,1)
free-lunch (0,0)
free-lunch (1,0)
free-lunch (0,1)
free-lunch (1,1)