New Directions in the Evaluation of Title I

Stephen W. Raudenbush

University of Michigan

January 31, 2002

This paper was the basis for the presentation on Optimal Designs for Evaluating Whole-School Interventions given at the American Educational Research Association annual meeting in New Orleans, April 2002.

In 1965, Congress passed the Elementary and Secondary Education Act and President Johnson signed it. A key provision of the act was Title I, which earmarked funds to improve the educational achievement of children living in poverty. Despite recurrent public debate over its utility, and despite changes in the legislation that supports it, Title I has endured over the past 37 years in various forms and now provides about 9.5 billion dollars annually in federal assistance to local schools (US Department of Education, 2001).

Debate over Title I has addressed two broad questions. First, does Title I make a substantial contribution to the achievement of poor students? And second, given a commitment to Title I funds, how can those funds be best used to support school achievement of these children?

Paralleling the debate over Title I is a recurring debate over how to evaluate the Program. The aim of this paper is not to revisit the history of Title I evaluation, but rather to consider new strategies and designs for evaluation. It is useful, however, to recognize that the two broad questions defined above have quite different implications for evaluation strategy.

If the key question is whether Title I makes a useful contribution to the learning of poor children, the corresponding evaluation strategy would logically investigate the average impact of Title I on the achievement of the target students. Let's call this the "Title I vs No Title I" comparison, because the focus is on the question of funding the program versus not funding it.

If, on the other hand, the question is *how best to use* Title I funds, the evaluation strategy likely aims to compare different versions of Title I. In contrast to the “Title I versus no Title” I focus, the evaluation would consider the causal impact of getting “Version A versus Version B” of Title I, where the two versions instantiate interesting alternative conceptions of how to boost the achievement of poor kids using Title I funds.

The choice of the evaluation question has enormous implications for the design of research on Title I. I consider these implications briefly in the next section. I claim that the first question (Title I versus no Title I) is difficult to answer and that answering it would provide comparatively few valuable lessons about "what works" in education.

In contrast, pursuing the second question (“Version A versus Version B) is a question social scientists can, in principle, answer with some confidence, and that doing so is likely to produce large benefits for learning how to improve schooling.

Given the choice of evaluation question, I then shall consider design options. I assess the potential utility of randomized experiments, quasi-experiments, and survey research. Issues of sample size and power emerge naturally from this discussion.

The paper is therefore organized as follows. In the next section I consider evaluation strategies associated with the question "How does receiving Title I compare with not receiving Title I?" Next, I consider evaluation strategies associated with the question "How do alternative

approaches to using Title I funds affect student learning?" I then consider implications for sample size and power before drawing final conclusions.

Evaluating the Overall Impact of Title I

If the key policy question concerns the overall impact of Title I, one would likely frame an evaluation study to assess how much the target children have gained in achievement, on average, as a result of Title I spending. The causal question, posed more sharply, would be to compare the current level of achievement of the typical Title I student to the level of achievement that this student *would have displayed in the absence of Title I*. Statisticians refer to this latter quantity as the "counterfactual outcome." We can't directly observe the achievement level that target children *would have attained* in the absence of Title I because we cannot turn back the hands of time and observe these kids in a "no-Title I" world. Instead, we would aim to compare the Title I children to *very similar* children who for some reason did not have access to Title I funds. Always a difficult, this task has become more difficult as Title I has become more widely institutionalized. One must ask how the "non-Title I" poor children escaped access to Title I funds and whether the characteristics associated with escaping Title I involvement might be linked to outcomes. Answers to such questions are highly speculative.

As the program has become more universally established, it has become more plausible that non-Title I children live in circumstances and attend schools that are really quite different from those of the children who *do* have access to Title I funds. This lack of comparability undermines the validity of causal inferences about the average impact of Title I.

The soundest way to answer a causal question is to conduct a randomized experiment. Classrooms, schools, or districts would be assigned at random to receive Title I funds or not to receive those funds. Random assignment would insure the comparability of those receiving or not receiving the benefits of the funding. It would be difficult, however, to imagine how such a randomized experiment could be conducted. It would be hard to find an ethical, political, or legal justification for depriving some eligible children of Title I funding.

In the absence of a randomized experiment, a clever researcher would likely try to find accidental reasons why one group of students did receive funding while a nearly identical group did not. For example, new census data become available every ten years, and at that time Title I funding streams are re-directed. Some schools that previously had no Title I funding become eligible as a result of demographic shifts revealed in the census and receive an influx of Title I funds. This sudden influx of funds can be viewed as creating a kind of natural experiment. Cohorts of children attending those schools *after* the influx of funds can be compared to cohorts of children attending those schools *before* the influx of funds. And the children enjoying the benefits of those funds can be compared to children in districts or schools that *just missed* eligibility for the influx of funds or to those who *suddenly lost* funding.

Natural experiments like that described above are often the best option for valid causal

inference in the absence of a "true experiment," that is, a study that formally randomly assigns units (kids or classes or schools) to "treatments." The generalizability of results from such a natural experiment is often limited, however. The study sketched in the previous paragraph would tell us nothing, for example, about the impact of Title I on kids in districts with poverty rates far above the cut-off point for Title I funding. Such districts are likely to be especially disadvantaged and therefore to be of special interest in any Title I evaluation.

An entirely different kind of criticism might be aimed at a study that sought to assess the average impact of Title I. Title I funds are used in myriad ways across districts. For this reason, Title I has often been termed a "funding stream" rather than a program. A funding stream can generate many "programs," that is, many more or less coherent strategies for using the funds to improve teaching and learning. Such programs might include "pull out programs," enrichment programs, inservice training efforts, or a whole array of "whole-school reform" efforts.

Suppose that the highly varied set of programs produced highly varied effects on students. It might well be true, in fact, that some program options boost student achievement substantially while others have negative effects. To assess the "average effect of Title I" would mask these highly varied effects. A finding that Title I has "a small positive effect," for example, might be true on average, but this finding would misrepresent the impact of Title I on the many students experiencing either very good or very bad implementations of the program.

Moreover, focusing on "the average effect of Title I" implies a comparison between Title I and some ill-defined control group. Such an ill-defined alternative to Title I will likely change over time so that a study of Title I's average effect in, say 1980, may have no relevance to Title I's average causal effect in 2000. In reality, both "Title I" and the alternative or "control" treatment are so ambiguous that little can be learned from such a study about how to improve instruction for poor children.

In summary, there are three big problems with a “Title-I versus no Title I” evaluation. First, it is hard to find a credible answer to this causal question because randomized experimentation is hard to justify and the near universal implementation of the program makes it difficult to find a good control group. Second, the treatment (“Title I”) and the control group (“no-Title I”) are ill defined, so that a comparison between the two gives us very little theoretical understanding of what works in education. Third, the meaning of both of these conditions is likely to shift with time and across locations, leading to large uncertainty about the generalizability of any such finding.

Comparing Versions of Title I

The current policy debate appears to focus more on *how* to do Title I rather than *whether* to do it. If so, a study of the average impact of Title I links poorly to the policy context. The more sensible evaluation focus concerns the effects of alternative approaches to Title I.

From a methodological standpoint, a study of alternative approaches has two appealing features: a) sound causal inferences become accessible; and b) the theoretical specificity of the study design makes it likely to produce important new insights about how to organize schools and conduct instruction.

Obtaining Causal Inferences About Versions of Title I

As mentioned, a randomized experiment provides, in principle, a sound basis for causal inference. If the aim of an evaluation is to assess the impact of alternative approaches to Title I, one can readily imagine a randomized experiment that is ethically sound and politically feasible. Let us consider a couple of examples.

Example 1: Assessing the Impact of Whole School Reform

Consider a popular program such as Success for All, which now is working in more than 1000 elementary schools in an attempt to boost early literacy (Slavin and Madden, in press). Many schools want to adopt the program but it is expensive and the resources available are limited. Indeed, it is impossible to simultaneously implement the program in every school that wants it.

One might seek schools to volunteer to get the program at no cost or a reduced cost. All volunteering schools would ultimately receive the program, but the timing -- that is, which schools get the program first -- would be decided by a lottery. A lottery is a perfectly fair way to decide this question, given that resources do not allow all interested schools to receive the program simultaneously. The schools assigned to receive the program immediately become the experimental group while the schools selected by lottery for delayed implementation become a randomized "wait list control group." The two groups are compared during the period before the wait-list control schools receive the program.

Two strategies make thus this kind of approach ethically sound and practically feasible: 1) the use of a wait-list control group; and b) the assignment of schools rather than kids to treatments.

Example 2: Comparing Two Whole-School Reform Strategies

A second kind of randomized study does not compare an experimental group (e.g., Success for All) to a control group (e.g., whatever a school would have adopted in the absence of Success for All). Rather, there may be two alternative programs -- both attractive -- that can be compared. If we really don't know which works better, we can argue for randomized experimentation, providing, of course, that participants are willing to try either approach. This latter condition may not hold, in which case a well-controlled but non-randomized study may be needed.

Yield for Educational Theory and Practice

As mentioned, a "Title I vs no-Title I" comparison has little potential to yield much new insight about teaching and learning. As a funding stream, "Title I" is, in reality a complex array of programs that shift with time while "no Title I" might include any number of approaches to school organization and instruction. Knowing the average causal effect of Title I versus no-Title I thus tells us very little about how to improve schooling.

Comparing a specific use of Title I funds to a control has somewhat more potential for producing new knowledge. Success for All, to choose just one example, is based on a well-specified theory of instruction and a coherent set of strategies for how to assess student learning and how to respond instructionally given the assessment. Nevertheless, the theoretical yield of a "Success-For-All versus control" comparison is constrained by the unspecified nature of the control condition. Assuming that instructional methods vary randomly across control sites, the impact of Success for all versus control comparison will also vary randomly. While some attempt may be made to capture information about methods of school organization and instruction at the control sites, an experiment of this type will likely be sub-optimal for learning about educational practice.

A study that compares two reasonably well-defined strategies for instructional improvement is likely to produce more useful new knowledge than the studies considered so far. This will be particularly true if the interventions are selected to produce maximally interesting contrasts between educational approaches. In particular, suppose schools are assigned to two well-defined interventions and that researchers carefully assess instructional practices at each participating school. The observed instructional differences between interventions are likely to be large relative to the variation within interventions. This between-intervention difference in measured instruction creates, in principle, an opportunity to examine in some detail the consequences of instructional practice for student learning.

Indeed, the theoretical usefulness of such a study may well exceed that of even the best designed large-scale sample survey. In large-scale surveys, instructional practices tend to covary with district and school context and with the demographic background and prior instructional history of the students. In a sense, instruction and context likely have achieved a kind of mutual accommodation or "equilibrium" in which the direction and magnitude of causal effects become difficult or impossible to disentangle.

There are three aspects of a comparison of interventions that create opportunities to learn about the causal effects of instruction, opportunities that do not arise in a survey. First, if schools are randomly assigned to programs or even if schools are matched with respect to context, the students who experience different types of instruction associated with the two programs are more likely to be comparable than in the case of a large-scale survey. Second, a study of well-designed interventions creates an opportunity to see what happens when agents of change deliberately intervene, disrupting the equilibrium that may have developed between instruction and context. Third, if the interventions are based on relatively well developed theory and are well implemented,

instructional alternatives are likely to appear in a more "pure form" than would arise in a survey. The sharpness of the contrast between these alternatives increases the chance of finding theoretically interesting effects (Rosenbaum, 1999).

The research strategy emerging from this discussion takes us beyond questions about "on average" effects of Title I or of particular programs. Instead, by comparing sharply different alternative approaches to teaching, the aim is to discover how big the effects of Title I *can be*. If the potential effects of Title I funds are found to be large in the presence of specific instructional interventions, then our research enterprise will prove extremely useful to those seeking effective methods of teaching poor children.

The search for powerful instructional interventions based on Title I funds may also begin to suggest upper bounds for the effects of Title I. Such upper bounds may be of great use in the longstanding debate over the funding levels for Title I. This link between research on instructional interventions and research on resources is the topic to which I now turn.

Instruction, Intervention, and Resources

Cohen, Raudenbush, and Ball (2002) sketch the history over the past four decades of research on educational resources and educational achievement. Citing limitations in this work, they call for a new stream of research that puts alternative instructional approaches in the foreground and uses research on instruction as a vehicle to examining effects of resources. I summarize the argument briefly here because of its relevance to Title I evaluation.

Research on effects of resources in education has typically followed an implicit or explicit causal model. In that model, the fundamental causal variable is some resource, for example, per-pupil spending, teacher-pupil ratio, class size, some measure of teacher qualification, or the availability of some sort of equipment (e.g., computers) or facilities (e.g., a science lab). The fundamental outcome is student achievement based on a test score. Some studies also include key moderator variables. A moderator is a third variable that determines the magnitude of the impact of the causal variable on the outcome. For example, the effect of per-pupil expenditures will depend on how the expenditures are allocated, e.g., whether the expenditures are used for academic and non-academic procurement. In this case, expenditures are the causal variable, achievement is the outcome, and use of the expenditure (academic versus non-academic) is the moderator. In a second example, availability of computers is the causal variable, achievement is the outcome, and use of computers is the moderator. The availability of computers has an effect on achievement only to the extent the computers are effectively used in instruction.

The research paradigm sketched above has had the useful effect of deflating optimism about the extent to which simply investing in conventional resources can be expected to boost achievement. With some notable exceptions, resource effects appear small. But this finding, while initially instructive for policy, doesn't take us very far. Indeed, we are left in a dilemma. On the one hand, we know that kids learn a lot in school. For example, Bryk and Raudenbush (1988)

compare summer and academic-year learning rates and find huge differences: kids learn at a much more rapid rate during the summer than during the academic year. We also know that education cannot occur without resources such as teachers, buildings, classrooms, and books. Given that schooling has a big effect and that schooling cannot occur without resources, a research tradition that tells us "resources don't matter" is met with healthy skepticism. Moreover, such a research paradigm tells us little about how to invest scarce resources.

The key flaw in the research paradigm, according to Cohen et al., is the causal model. Resources do not "cause" student achievement in any proximal sense. Instead, the effect of injecting a new resource into the system is contingent on an extremely large and interacting set of moderators. In this setting it is extremely difficult to trace the contingencies that determine the effect of the resource. For this reason, resources should not be the key causal variable in the model. Instead, the authors argue that instructional practice and its causal link to student learning should be at the center of the model. In particular, more or less coherent alternative approaches to instruction in the setting of classrooms are the essential proximal causes of student learning. Of course any reasonably well-specified notion of instruction (which the authors call an instructional "regime") entails the use of resources. Thus some level of resources must be available. The level of resources required to achieve an outcome under a given regime is, however, an empirical issue.

This leads to an alternative causal model in which the causal variable is instruction, the outcome is learning, and resources are moderators. As mentioned, a moderator is a third variable that determines the magnitude of the effect of the causal variable on the outcome. Resources are moderators because constraining resources, in principle, moderates the impact of a good instructional approach on student learning. For example, reducing skill requirements of teachers may undermine the impact of an innovative writing program on student learning.

In sum, the standard research paradigm for studying resources views resources as the cause, learning as the outcome, and resource use as the moderator. The big problem with this model is that instruction is absent from it. Instruction is by all accounts the essential proximal cause of academic learning that occurs in a school setting, so to ignore instruction is hardly a minor oversight. The alternative model makes instruction the key cause and resources become the moderators.

This alternative causal model leads to a two-part research program to ascertain the importance of conventional resources for student learning.

The first step is to identify instructional regimes and assess their causal effects on student learning. An instructional regime is a set of more or less explicit rules for choosing how to teach given the current level of student understanding. Thus, if instruction involves the interplay between teachers and students around materials, an instructional regime is the programmatic aspect that operates independently of the students taught. The set of rules concerning how to treat student responses exists independent of the students or how they respond.

Having identified truly effective regimes, step two is to investigate the level of resources required to achieve their positive effects. A new approach to math instruction, demonstrated in a study of teachers with advanced degrees in math, may be found to achieve nearly as good results when high-quality in-service training is provided teachers who do not have such advanced pre-service training. A program of writing instruction may work nearly as well in a classroom of 20 as it does in a classroom of 15.

This two-step paradigm is potentially relevant to the evaluation of Title I. Some whole-school reform programs are quite expensive and their cost will tend to limit their adoption. If we knew how effective these programs *can* be, in the presence of generous resources, we might then investigate how constraining resources constrains their effects. Thus, resource effects are conceived in relation to instructional aims and means.

This alternative paradigm has parallels in medical research. Medical research is not organized around discovering the effect of per-patient spending on health, nor do we view alternative drugs as moderating links in the causal pathway between spending and health. Instead, we ask, for example, whether heart bypass surgery works better than medication in promoting the health of angina patients. We then seek the most cost effective ways of delivering the desired treatment.

The engine that drives this paradigm for research in education, then, is the causal comparative study of alternative instructional regimes. If Title I funds are to affect student learning, they will do so largely by supporting more or less coherent approaches to instruction. To understand the possible impact of those funds requires, then, carefully controlled studies of instruction, tethered to research on how instruction fares under resource constraints.

The Role of Surveys and Ethnographies

The research program sketched so far is an activist program. Rather than passively observing current conditions, it puts interventions into place in order to test their effects, and deliberately varies resources to gauge their moderating influence on instructional effects.

Yet more passive forms of inquiry, including surveys (but also including ethnographic studies) also have useful roles within this framework. Surveys can tell us how poor children are faring nationally, how large gaps are between these children and more advantaged children, and whether those gaps are growing. Such research can tell us about rates of learning and whether these growth rates are different for different kinds of kids in different settings. Surveys can tell us the range of instructional approaches Title I funds support. Ethnographies can suggest problems that program designers must attend to and new strategies for teaching. They can tell us how teachers and children interpret their experiences in instruction and enable us to test the validity of our quantitative assessments of teaching and learning. Information from surveys and ethnographies can suggest new interventions and new lines of intervention research.

While multiple approaches to research are essential, given the magnitude of the Title I enterprise, I argue here that the balance of research effort now needs to be titled toward assessing the causal impact of instruction supported by Title I funds. A more activist research agenda will tell us a great deal about how Title I funds should be used and how much improvement in student learning we might reasonably expect from such a funding stream.

Design Issues in Intervention Research on Title I

The discussion so far emphasizes the importance of studies that assess the causal effects of instructional approaches supported by Title I. Such studies might be randomized experiments or they might be carefully controlled non-experimental studies (sometimes called quasi-experiments or observational studies). They might be cross-sectional or longitudinal. Given these choices, sample sizes must be determined, including the frequency of observation and duration of a longitudinal study, the number of children sampled per school, and the number of schools sampled. A detailed consideration of these issues depends on the specific aims and strategies selected for a given study (Raudenbush, 1997; Raudenbush and Liu, 2000, 2001). However, I can briefly review some of the general concerns that are likely to cut across studies.

Randomized Studies

Design issues concern the unit of random assignment, the duration of the study, and sample sizes.

Unit of assignment. Randomized studies of whole school reform will naturally use the school as unit (Cook et al., 1999a, 1999b). The school is the unit to be treated, so no lower levels of assignment (e.g., classroom or child) make sense. In fact, a thought experiment reveals that a study that would assign students at random to schools would not answer the causal question. Such a study might show that one set of schools produced higher achievement than did another set of schools, but such a difference could not be equated to the causal effect of the reform. The reason is that the key form of selection bias is at the school level. Certain schools with certain kinds of leadership and teachers are more likely to seek out a whole-school reform intervention than are other schools. To overcome such a bias requires random assignment at the level of the school.

For instructional interventions that operate on individual classes rather than the whole school, one might imagine randomly assigning classes rather than schools to interventions. However, such an approach may prove unworkable for several reasons. First, a "diffusion of treatments" is likely as teachers in one intervention program share their new methods with teachers in an alternative intervention or a control group. Second, assignment of classes at random must be repeated each year to maintain clarity about causal effects over time. This would require teachers to be ready to switch interventions each year, an implausible prospect. Thus, studies of instructional interventions using the class as the unit of randomization are most

plausibly limited to those that can be evaluated in the course of a single academic year.

When schools are assigned at random to interventions, however, studies of instructional practice are non-experimental. There will be variation within interventions on instructional practice but these differences between classrooms will depend on the teacher and student background. Thus, while causal inferences about the differences between interventions are protected by randomization, causal inferences concerning differences in instruction within interventions will not be so protected.

Longitudinal versus cross-sectional design. A design that follows students over time after the introduction of the treatment is attractive for several reasons. First, one can assess the impact of the intervention on the rate of learning as well as on the status of the student. Studies of learning rates may lend more statistical power to estimates of causal effects because variation within interventions on these rates is likely to be smaller than variation within interventions in status. Second, a longitudinal design allows an assessment of longer-term consequences of the intervention. It would be interesting to know whether intervention effects are sustained or even accelerate or whether they fade as the child's school career proceeds.

There are risks, however, in adopting a longitudinal design. The greatest risk comes from potential bias associated with differential attrition. Mobility rates are often high in urban school districts. Imagine an intervention so effective that parents of children who are faring poorly make great efforts to enroll their children in the intervention school. The in-migration of such children would tend to pull down the mean of the intervention school, biasing the estimate of the causal impact. Now one might decide not to take such in-movers into the sample. However, rates of out-migration -- and reasons for out-migration -- might also vary across interventions. Perhaps parents with the means to move from a poor neighborhood decide not to because the local school is participating in an effective intervention. This would likely bias the intervention effect upward, since such compartively high-income children would be expected to display above-average achievement.

Care must also be taken in terms of how students retained in grade and special education students are handled. An effective intervention might prevent certain students from grade retention or assignment to special education. This would produce a bias unless all children, including those retained and those in special education, are assessed.

One might conclude from this discussion that a short-term cross-sectional evaluation would be safer than a longitudinal study in a randomized setting. One shortcoming of a cross-sectional study, already mentioned, is its inability to trace long-term effects. A second is that interventions take time to mature and become fully implemented. So a short-term study may miss the impact of the intervention.

In sum, differential attrition and non-random missing data convert an initially randomized experiment into a non-randomized or "quasi-experiment." Moreover, as mentioned, studies of

instructional differences within interventions are non-experimental. These facts may encourage some restraint about the utility of randomized experiments. However, non-randomized studies are subject to these difficulties as well and to other difficulties as well.

Non-Randomized Studies

In a non-randomized study, one hopes for large, theoretically interesting differences between interventions in their educational approaches and small differences between them in terms of the types of students and teachers they serve (Rosenbaum, 1999). One might imagine a district in which all schools with achievement means below a certain level are required to select from several very different interventions. Because the schools are in the same district and because all have low achievement, their students may be similar. And because the interventions are different, the educational effects under investigation may be large. But one must be on guard. Interventions may develop reputations for serving particular kinds of students or teachers, and this reputation may encourage some schools and not others from enrolling in the intervention. This may lead to substantial pre-treatment differences between interventions in the school settings and types of kids they serve. And sometimes interventions look more different on paper than they do in the classroom.

Matching by district and by prior achievement is one way to approximate the comparability that would be insured by randomization. Another way is to use multivariate matching via the estimated propensity score (Rosenbaum and Rubin, 1983). The propensity score is the conditional probability of treatment assignment given a set of covariates, that is, background variables. Using a large-scale survey data base, it may be possible to match a small set of schools using a specific intervention to a very similar small set of control schools. The control schools are selected from the large reservoir of all schools because each of them has a propensity score that is nearly identical the propensity score of one of the "treatment" schools. It is a remarkable fact that matching on the propensity score insures that control and experimental schools will have similar distributions on all observed covariates.

While matching can be very useful, perhaps the most common approach to controlling potential confounding variables (pre-treatment covariates associated with treatment group membership) is through statistical control. The most common method is regression analysis. The observed achievement of a school in intervention A is compared to its predicted value if it had been assigned to intervention B. The predicted value depends on the validity of a linear model that uses data from schools in intervention B to predict how each school in A would do if it had had intervention B. Unfortunately, such model-based predictions are problematic when differences between schools, teachers, or students in the two treatments are large. Statistical control is best used in conjunction with matching rather than as a substitute.

A potentially strong quasi-experimental design uses annual testing data on students that is available from historical records. Consider, for example, data routinely collected in a state with an annual testing program. One might be able to use such data to estimate growth trajectories for

students attending a given school before and after the implementation of an instructional intervention. The pre-treatment growth data serve as controls for the post-treatment growth data. If, in addition, such growth data are available from other schools that never receive the treatment (or that receive an alternative treatment), a second form of control is available. Such designs may be called multiple cohort designs and can provide a comparatively strong basis for causal inference in the absence of randomization. In essence, the combination of two types of information – data from earlier cohorts in the same school and data from other schools – is combined to predict how well intervention kids would have done without the intervention.

Sample Sizes

The discussion so far has focused on strategies for reducing bias by means of random assignment, matching, or other means. A second important issue concerns the precision of estimated treatment effects and the power of tests of significance of those differences. Statistical power depends on the magnitude of the effect under investigation ("effect size"), the sample size, and certain aspects of the design.

Effect size. One hopes to compare interventions that are likely to have very different effects under one or more instructional theories of interest. For example, a theory of reading instruction that emphasizes phonics predicts that a phonics-intensive approach will work quite dramatically better than an approach that de-emphasizes phonics, even though critics of phonics might prefer the second regime. If the predicted difference between interventions according to all theories is small, hope diminishes that interesting results will emerge. Effect sizes are often described on a standardized scale. A standardized effect size is the mean difference between two post-treatment means divided by the standard deviation of the outcome. A standardized effect size of .20 to .30 is often regarded as large enough to be worth detecting, while effect sizes from .50 to .80 are often regarded as quite large in educational policy research.

Sample size. Many introductory statistics texts consider examples in which power depends strongly on the number of students per treatment. However, when schools rather than students are assigned to treatments, there are two sample sizes: the number of students in a school and the number of schools. Power for a cross-sectional comparison in this setting depends more strongly on the number of schools than on the number of students per school.

Cross-sectional versus longitudinal design. In a longitudinal design, power depends on the frequency of observation of the outcome, the duration of the study, the number of students, and the number of schools. Methods for deciding on frequency and observation are provided by Raudenbush and Liu (2002). For the purposes of this paper, we shall concentrate on cross-sectional designs while commenting on longitudinal designs.

Intra-school correlation. If differences among schools within an intervention are large, it becomes important to sample a relatively large number of schools per intervention in order to obtain a good estimate of each intervention's mean. It may, however, be much more expensive to

sample an additional school than to sample an additional student once a school has been selected. Raudenbush (1997) shows how to use information on the relative cost of sampling together with other information to determine the optimal sample size per school. Suppose it is the case that sampling schools is expensive. Then one tends to sample comparatively fewer schools and more students per school. Such a decision is especially sensible when the variation between schools is small relative the variation within schools.

Variation between schools (relative to the total variation) is indexed by a parameter called the "intra-cluster correlation." In our case, with schools as clusters, we shall label this the "intra-school correlation." The intra-school correlation is the fraction of variation in the outcome that lies between schools. If every school is a just a random sample of students, so that no variation lies between schools, the intra-school correlation is zero. At the other extreme, if all students within a school are identical but the schools vary, the intra-school correlation is 1.0. Intra-school correlations between .05 and .15 are common in US data sets.

An example. Statistical power in our case is the probability of rejecting the null hypothesis that the post-treatment population mean difference between two interventions is zero. It will increase with effect size and with the sample sizes (the number of schools and the number of students) and decrease as a function of the intra-cluster correlation for any fixed sample size.

Figure 1 displays the statistical power of a hypothetical intervention as a function of effect size (set at .30 or .50), and the intra-school correlation (set at .05, .10, or .15). The critical significance level is set to be .05. Sample size per school is held constant at 50 and the number of schools is allowed to range from 10 to 100. One sees that power increases with effect size (note that when the effect size is 0.50, power is larger than when the effect size is 0.30 for every intra-school correlation). Also, power is greater when the intra-school correlation is smaller. In the worst case scenario (small effect size of 0.50 and large intra-school correlation of .15), power exceeds .80 when the total number of clusters is 60 (30 per treatment). For any other combination of effect size and intra-school correlation, power exceeds .80 whenever the total number of schools is about 40 or larger.

Figure 2 displays statistical power under the same effect sizes and intra-school correlations. Now, however, we hold the number of schools constant at J = 50, while allowing the number of students per school to vary from n=10 to n=100. One sees that power is comparatively insensitive to n. Once n reaches about 40, power changes little by adding more students per school. At some point massive investments in n, holding J constant, give little added benefit in terms of power.

Longitudinal data. As mentioned earlier, planning for adequate power is more complex when one is comparing the growth curves of students in two interventions rather than comparing their means. However, in previous work I have studied this planning problem in some detail. Assuming a sample size of 75 students per school with interest in following each student from kindergarten through grade 5, I found a choice of 25 schools per intervention to be adequate to

detect modest treatment effects on growth. I used data from the Prospects evaluation of Title I to estimate the variation in growth rates between and within schools for this purpose.

Power for assessing teacher effects. Recall that comparisons of teachers' instructional approaches within interventions are non-experimental unless teachers are assigned at random to treatments within schools, which is implausible. Thus, one must attempt to identify and control bias in estimating these effects. Power, however, tends to be larger than for comparing schools. Thus, if power is adequate to learn about mean differences between interventions, it will also be adequate to study differences in teacher effects.

Final Remarks

In sum, we have considered statistical power in a cross-sectional study that assigns schools to treatments (either randomly or not). If the data are similar to those collected in the Title I evaluation known as Prospects, we find that a total sample of 60 schools, equally divided between two alternative program types, with about 50 students per school, is large enough to detect an effect size of .30 in standard deviation units even if the variation between schools (within treatment groups) is larger than expected. A sample of the same size will likely have more power to detect student differences in growth in a longitudinal context or a non-experimental comparison between different types of teachers within schools. The needed sample size will certainly vary from study to study for a variety of reasons, but the results here give some rough guidance to policymakers and planners thinking about evaluating Title I by comparing school-wide instructional reform efforts. The software that produced these graphs is available without charge from the author and is easy to use.

The strategy of comparing alternative instructional interventions as a way of evaluating Title I has considerable appeal. If the studies are thoughtfully designed, they ought to provide a sound basis for causal inferences about the consequences of using Title I funds in theoretically interesting ways. A series of such studies is likely to produce a high yield of new understanding about “what works” instructionally for poor children and how Title I funds might best be used to achieve its goals.

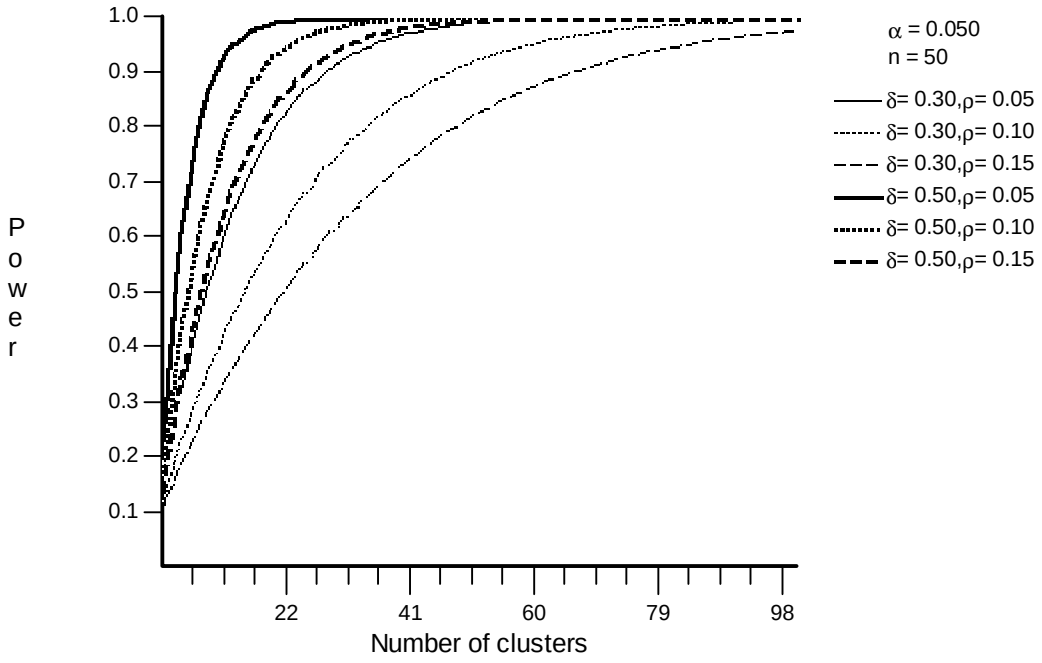

Figure 1: Power as a function of J (the number of schools) at various effect sizes and intra-school correlations), holding the within school sample size constant at n = 50.

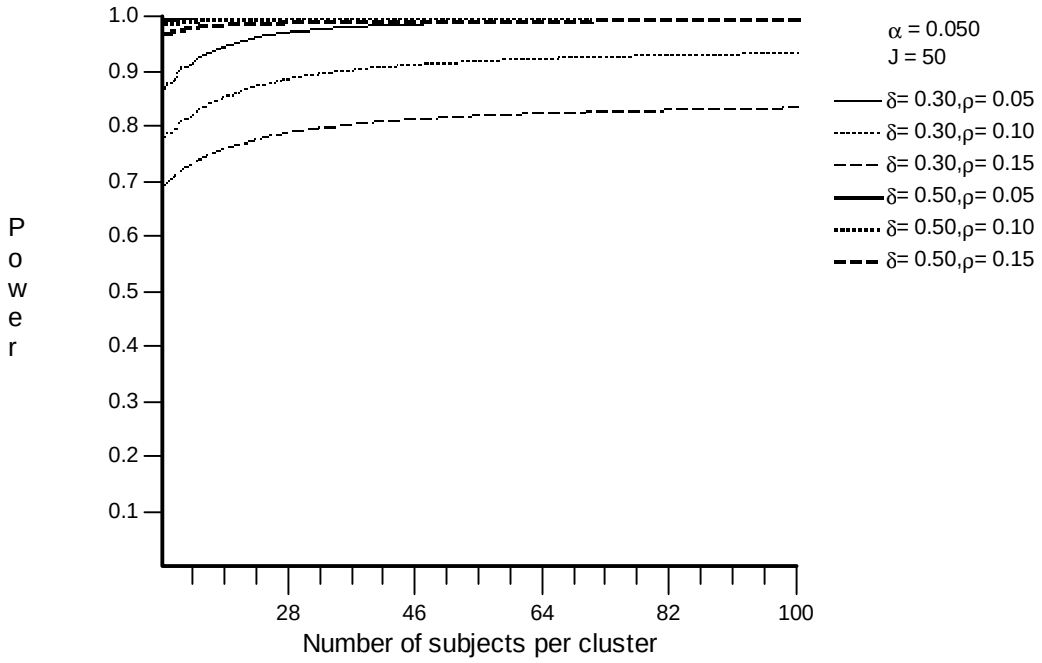

Figure 2: Power as a function of n (the number of children per school) at various effect sizes and intra-school correlations), holding the number of schools at J = 50.

References

Bryk, A.S., & Raudenbush, S.W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. American Journal of Education, 97, 1, 65-108.

Cohen, D., K., Raudenbush, S. W., & Ball, D. L. (in press). Resources, Instruction, and Research.To appear in Boruch, R. and Mosteller, F., *Education, Evaluation, and Randomized Trials*.Washington, DC: The Brookings Institute.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999a). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal, 36*(3), 543-598.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (1999b). *Comer's school development program in Chicago: A theory-based evaluation.*, Northwestern University.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. Psychological Methods, 2(2), 173-185.

Raudenbush, S.W. & Liu, Xiaofeng. (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods, 5(3), 199-213.

Raudenbush, S.W. & Liu, X. (2001). Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change. Psychological Methods, 6(4), 387.401.

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

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

Slavin, R., & Madden, N. (In press). *One million children: Success for all.* Thousand Oaks, CA: Corwin.

US Department of Education (2001). *High Standards for All Students: A Report from the National Assessment of Title I on Progress and Challenges Since the 1994 Reauthorization*. Washington, DC: Office of the Undersecretary, Program and Planning Service, US Department of Education.