Designing Field Trials of Educational Innovations

by Stephen W. Raudenbush
University of Michigan

October 24, 2003

This paper was prepared for a national invitational conference "*Conceptualizing Scale-UP: Multidisciplinary Perspectives*," to be held November 3-4, 2003 in Washington D.C. The work reported here was supported by a grant from the W.T. Grant Foundation ("Building Capacity in Evaluating Group-Level Interventions," Stephen W. Raudenbush and Howard Bloom, co-principal investigators).

Abstract

The task of interest here is to establish compelling empirical evidence of an educational innovation's impact when enacted across a broad array of field settings. By assumption, the intervention affects student learning by changing instruction. Of interest is the impact when resources are abundant as well as the sensitivity of the impact to resource constraints. In the ideal study, agents (tutors, teachers, or counselors) or settings (classrooms, schools districts), rather than students, are randomly assigned to the innovation or to some well-defined alternative experience. Non-experimental alternatives are approximations to this ideal. Statistical precision depends more on the number of settings than on the number of students, and can be enhanced by pre-treatment blocking, use of covariates, or longitudinal designs. Causal generalizations about the range of possible effects arise primarily from syntheses of streams of studies that vary not only in the availability of resources, but also by the instantiations of the innovation and the alternatives, by the backgrounds of the agents and participants, and by the organizational arrangements that facilitate or constrain implementation.

Let us distinguish between two related problems. The first arises when innovators take novel educational practices "to scale." These practices may involve new instructional methods, new approaches to school organization, or new applications of computer technology. Typically, the innovative practices will have produced encouraging results in one or more local settings. In taking the innovation "to scale," the challenge is to implement these practices in a much larger range of settings without sacrificing the qualities that made the innovation appealing in its original venue. Characteristic difficulties arise in exporting the innovation from its setting of origin to its settings of destination: The inventor of the innovation will have less control over implementation in the new settings than in the original setting; teachers in the new settings may be less enthusiastic, resources supporting the innovation may be less abundant, and organizational conditions constraining the innovation more pronounced in the new settings than in the original setting. Moreover, effective implementation may require that the innovation be adapted to the specific new circumstances, including, for example, the organizational routines, teacher skill levels, available resources in the new setting. If so, essential ingredients for success present in the original setting may no longer characterize the innovation as it is enacted in the new settings.

The second problem is to establish compelling empirical evidence of the innovation's impact across a broad array of new settings. In many cases, convincing evidence will exist that the program had positive effects in its original local setting. Unfortunately, the difficulties mentioned above that commonly arise in taking the innovation to scale will tend to diminish the impact of the innovation in the new settings. Its effects may also be more variable across the more diverse new settings. And experimental control may be more difficult to achieve in the "field" (the new settings) than in the "lab" (the setting of origin).

In this paper, I consider the second problem and its implications for research design. Specifically, I ask how evaluations should proceed when innovations are taken to scale. Based on a simple conceptual framework, I ask how studies might be designed to minimize bias, maximize precision, and facilitate causal generalization. I close by considering some implications for improving the educational evaluation.

Conceptual Framework

I confine my interest to innovations that aim to improve student learning. Following Cohen, Raudenbush, and Ball (2002), I define instruction broadly to include teacher and student interactions in classrooms around materials and reason that the innovations of interest must influence instruction if they are to affect academic learning. This assumption is depicted in the causal model displayed in Figure 1, which displays causal link (1) between the innovation and instruction and a causal link (2) between instruction and student learning. The dashed line (3) implies that there is typically no direct effect of the innovation and student learning, that is, no effect of the innovation on learning that does not work through instruction.[1]

[1]Certainly we can imagine exceptions. Consider a home visiting program designed to improve

Insert Figure I About Here

Interplay between resources and innovation. A key feature of the conceptual model is that the impact of the innovation on instruction is moderated by (4) the availability of resources. Teacher skill and commitment and class size; students' nutrition, mental health, prior knowledge, and academic motivation; and administrative effectiveness, among others, are resources that may enhance or constrain the effect of a new innovation on the quality of instruction, and, hence, on student learning.

Before a new innovation is "taken to scale," considerable prior research on its effectiveness will often be available. Consider for example, the Detroit Public Schools' recent district-wide adoption of the "Open Court" reading program in the early elementary grades. That program is based on a large and growing body of research concerning link 2 of our causal model, in this case the relationship between intensive and explicit phonetic instruction and student skill in sight-word recognition and reading fluency during the early grades (Foorman, Francis, Fletcher and Lynn, 1996; Snow, Burns and Griffin, 1998). Moreover, let us assume that this causal connection has been established for children who are demographically similar to those who attend school in Detroit. Then, it is reasonable to hypothesize that if adoption of Open Court significantly enhances that kind of explicit instruction (link 1), Detroit children will benefit, at least in terms of sight-word recognition and fluency. However, the local settings in which prior research established the links of interest may be quite different from the Detroit settings in which the innovation has recently been adopted. Specifically, the resources available for effective implementation of the program may be considerably more scarce in Detroit than in the settings for the original research. Thus, the first (link 1) between the innovation and instruction may be weaker than in past research.

Now consider a hypothetical, well-designed study of the impact of Open Court on student learning in Detroit. Suppose that no effect were found, not because explicit instruction is ineffective in Detroit but because resource constraints in Detroit precluded effective implementation so that no substantial increase in direct instruction resulted from adoption of Open Court. Such a result would be impossible to interpret if this hypothetical study did not also assess the impact of the innovation on instruction.

Past research on program evaluation has distinguished between "theory failure" and

---

parent encouragement and student motivation. The genesis of the effect is in the home, but if the newly developed motivation failed to achieve expression in the classroom, it would likely have little long-term effect on learning. Direct effects of innovations on learning are thus likely to be weak at best.

"implementation failure." Theory failure occurs when a program is well implemented but does not have the impact hypothesized by the program theory. Implementation failure occurs when a program is not effectively implemented, in which case the program theory was never tested. In this case, a finding of "no effect" reflects implementation failure not theory failure.

A two-stage strategy. In the model I propose, implementation -- that is, the impact of the innovation on instruction -- depends on available resources. An ideal stream of evaluation research would have two stages. In the first stage, research would assess the impact of the innovation under extremely favorable conditions. Implementation would be carefully assured, staff development would be generously funded, teacher skill levels would be high. The purpose of this stage of the research is to demonstrate that the innovation can, under the most favorable conditions, promote a substantial effect on student learning. A failure to do so would then give evidence of theory failure, not implementation failure.

Having demonstrated a truly effective innovation, the next stage of the research would investigate the level of resources required to produce a strong effect. A new approach to math instruction, tested initially in a study using teachers with advanced degrees in math, may be found to achieve nearly as good results when high-quality in-service training is provided to teachers who do not have such abundant pre-service education. A program of writing instruction may work nearly as well in a classroom of 20 as it does in a classroom of 15.

Such a two-stage strategy is often employed in clinical trials in medicine that distinguish between treatment efficacy and treatment effectiveness. Treatment *efficacy* is the magnitude of the treatment effect under optimal conditions. In contrast, *effectiveness* is the magnitude of the treatment effect under resource constraints operating in a field setting. Efficacy studies are important because they demonstrate the a new treatment can substantially reduce the impact of a disease. This spurs effectiveness studies -- studies of attempts to make the new treatment widely accessible.

A useful example involves reading comprehension. At a recent meeting of the Technical Working Group of the Title I Review Panel, reading experts expressed agreement that cognitive science has demonstrated in controlled settings that certain specific instructional techniques can enhance reading comprehension. What has not been shown is that elements of these can be combined within the context of the regular school day in order to produce big effects on the comprehension of children attending high-poverty schools. Because the problem of reading comprehension for disadvantaged children is so enduring and so central to educational improvement, the experts agreed that it would be extremely important for policy to demonstrate that an intervention can have a large effect -- even if the resources available in that demonstration were unrealistically abundant. Indeed, an "existence proof" of such an effect would spur considerable public enthusiasm for a stream of studies designed to test the replicability of the effect when resources are more realistically constrained.

What is the comparison group? Now consider a hypothetical situation in which Open

Court had a much bigger effect in Detroit than in prior research, not because it was better implemented there, but because the comparison treatment was worse than in prior research. Unless the researchers had documented the quality of instruction in the comparison classrooms, the cause of the enhanced effect in Detroit would be a mystery. We consider the definition of the causal variable later (see "causal generalization").

Treatment-by-subject interactions. So far we have assumed that prior research on Open Court was based on children demographically similar to those in Detroit. If this is not true, or if the Detroit children vary in other ways that moderate the effect of the treatment, differences between the effect found in Detroit and the effects found earlier may remain mysterious. To clarify these differences requires a plan to specify child background characteristics that moderate the effect of the treatment.

Summary. In sum, a causal comparative study of an educational innovation in a field setting typically assesses the impact of that innovation on student learning. However, the effect will typically depend on how the innovation is implemented instructionally, which, in turn, will depend on resource availability. It will also depend on the quality of the comparison treatment and may depend on child background. Unless resource availability, quality of implementation, the nature of the comparison treatment, and child background are adequately specified, the impact evaluation will typically be difficult to interpret even if other aspects of study design are sound. Moreover, if treatment effects vary as a function of resources, the nature of the comparison treatment, and the background of the learners, a single study would not likely be sufficient to illuminate the "scale up" process. Rather, a series of studies may be necessary to clarify these moderating influences. Often a two-stage strategy is compelling. In this strategy, the first order of business is to establish an upper bound on the effect of an innovation. Subsequent work examines the dependence of the effect of resource constraints that arise in practical settings. In the section below under "causal generalization" we take up the question of how to combine results across a stream of related studies in order summarize the influence of "scope conditions" (resource availability, student background, comparison groups) on estimates of the effect of an innovation.

Minimizing Bias

A first-order policy question facing evaluators of "scaled up" innovations is to estimate, in a single study, the average impact of the innovation on student learning. Equally important for policy, as discussed above, is the dependence of that effect on the availability of resources. Resource availability affects the cost of the program but also specifies the contingencies planners must consider as they seek to produce the desired effect.

But what do we mean by "impact?" And how do we reduce bias in estimating impact? Comparatively recent thinking in statistical science defines the "impact" or "causal effect" of a program as the difference between the outcome observed for a student who received the program and the outcome that *this same student would have received* under an alternative educational

treatment (Rubin, 1978; Rosenthal and Rubin, 1983; Holland, 1986). The latter is the "*counter-factual*" outcome, which, together with the observed outcome, constitute the *potential outcomes* for the child in question. Of course, we cannot observe the counter-factual outcome because we cannot turn back the hands of time and observe that child under an alternative treatment condition. Thus, we cannot estimate the effect of the program on each individual child. Instead, we aim to estimate the *average effect* of the program on a sub-population by comparing the children receiving the innovation to very similar children who for some reason did not receive the innovation. The average difference between outcomes of these two groups is an unbiased estimate of the average causal effect of the program -- if treatment group assignment is "ignorable." Assignment is ignorable if the potential outcomes of the children are not statistically related to treatment group assignment.

For illustration, consider an evaluation that will compare children receiving Open Court to children in a control group who will receive the standard reading program. In this setting each child has two potential outcomes: an outcome that would be observed if the child experienced Open Court and the outcome that same child would receive if the child experienced the standard reading program. Now suppose that these potential outcomes are lower, on average, for children from low-income families than for children from high income families, and that low-income children are more likely than high-income children to be assigned to Open Court. Then an evaluation that ignored family income would be biased against Open Court. Here family income is a confounding variable -- a child characteristic related to the potential outcomes that also predicts treatment group membership. Treatment assignment is non-ignorable because the potential outcomes predict treatment group assignment.

Now suppose that the evaluator cleverly anticipates the confounding effect of family income. This researcher stratifies the sample by levels of family income and compares Open Court to the control group within these strata. This researcher has therefore eliminated the confounding effect of family income and therefore eliminated this source of bias. If family income is the only confounding variable, treatment assignment *within levels of family income* will be ignorable. The problem is that a host of confounders may exist, and unless the researcher anticipates and controls for all of them, treatment assignment will be non-ignorable and some bias in estimating the treatment effect can be anticipated.

Randomized trials. Suppose now that children were randomly assigned to tutors using either Open Court or the control program. Randomization of children would insure that no child characteristic is related to treatment group assignment. This would seem to insure ignorable treatment assignment and therefore eliminate bias. However, such a design would not eliminate confounding because it remains possible that more (or less) able tutors are assigned to Open Court than to the control program. Thus it is essential that *the tutors* be assigned at random to treatments. In fact, a thought experiment reveals that the random assignment of tutors to treatments is sufficient to eliminate bias in the evaluation. For example, even if the most able children are assigned to the best tutors, random assignment of tutors will eliminate bias in evaluating the impact of Open Court. Of course the results of this experiment would generalize

only to a world where good students get good tutors. Such a world may or may not be more realistic than a world in which students are randomly assigned to tutors. The key point is that a randomized design can, in principle, eliminate bias in estimating causal effects. However, it may not be sufficient or even useful to randomize at the level of the individual child.

Unit of randomization. Nearly all instructional innovations are implemented by agents in social settings. The lesson of the above example generalizes: the random assignment of students to agents or settings is generally insufficient to achieve ignorable treatment assignment and therefore to reduce bias. Rather, the crucial element in designing randomized experiments in the context of instruction is the *random assignment of the agents or the settings* to treatment conditions.

Consider another example: the evaluation of whole-school reform programs such as Accelerated Schools. The hallmark of the Accelerated Schools program is that it mobilizes the entire school staff to coordinate instruction to achieve common aims. Clearly the unit of treatment is the whole school. Now suppose an evaluator designed an experiment in which students were assigned at random to one of two sets of schools: a set of schools using the Accelerated School Model versus a set of schools without Accelerated Schools. After some appropriate period, the outcomes of the two groups of students would then be compared. Such a design would provide an unbiased estimate of the average causal effect of attending the first set of schools versus the second set of schools, but this effect could not be attributed to the Accelerated Schools Program.

The problem with the design employing random assignment of students is that it ignores the key selection processes, operating at the school level. It may be, for example, that only the most dedicated principals and teachers choose to adopt Accelerated Schools. Or perhaps it is the most troubled schools that seek help from such a whole-school reform effort. Either way, the pre-existing characteristics of the school leadership and staff, rather than the Accelerated Schools program itself, may be responsible for creating the observed differences between the two randomly assigned sets of students.

So the key to eliminating bias in this case is to employ random assignment of schools. Would random assignment of students help in any way? Such a step would likely reduce variation within treatment conditions, thus increasing precision (see the next section). However, randomly assigning students would also create an artificial world in which schools serve random subsets of kids. This is not the world we live in and therefore not the world in which the results of the evaluation would be applied or, in current jargon, "taken to scale."

Are randomized studies of educational innovations possible? Many educators have argued that "education is not medicine" and that randomized experiments in education are unfeasible or unethical or both. However, in my experience, closer scrutiny reveals that such opponents of randomization are imagining studies in which students are randomly assigned to educational treatments. Such studies may indeed be difficult to pull off, but as we have seen,

they are generally not sufficient to eliminate bias in any case. Rather, it is in most cases the random assignment of agents (tutors or teachers), or of settings (classrooms, schools, districts) that is essential to insure ignorable treatment assignment in studies of instructional innovation.

Thus, we see successful randomized studies of James Comer's whole-school reform program (Cook et al., 1999; Cook, Hunt, and Murphy, 1999) and an ongoing randomized study of Success for All (Slavin and Madden, in press). In these studies, schools were assigned at random to treatments. Such an approach is feasible and ethical when staff in many schools want to adopt a new program and the available resources are limited. Indeed, it is often impossible to simultaneously implement the program in every school that wants it. One might then seek schools to volunteer to get the program at no cost or at a reduced cost. All volunteering schools would ultimately receive program but the timing would be decided by a lottery. A lottery is a perfectly fair way to decide this question, and the schools waiting to receive the program become the control group for comparison with the treatment schools during the weighting period.

Robert Slavin and colleagues used an alternative ingenious strategy. Schools were assigned to receive Success for All starting in first grade or in third grade. Schools receiving the treatment starting in first grade are providing third-grade data control-group data for comparison with schools receiving the treatment in third grade. Similar, the latter schools provide first-grade control data for schools receiving the treatment in first grade. Ultimately all children in all schools will experience the program, but the staged implementation fits with program philosophy and resource constraints. Using this strategy, the researchers found it far easier to recruit schools to the experiment than by using the wait-list control method.

Are there alternatives to randomization? The beauty of randomization is that random assignment eliminates all confounding variables, including those the investigator might never have anticipated, insuring ignorable treatment assignment. However, randomized studies can be very expensive. As the next section shows, precision in school-based randomized studies depends heavily on the number of schools and recruiting and sustaining involvement of a large number of schools drives up the cost of the research. Nor is randomized experiment always possible.

When random assignment is not possible, a clever researcher will likely try to find accidental reasons why one group of districts, schools, or classrooms did receive an innovative program while another did not. For example, the timing of new legislation requiring schools to show improvement might vary across states, creating variation in the timing of implementation of an innovative program. Such a "natural experiment" might be viewed as an approximation to a randomized experiment. Conventional statistical strategies in econometrics, including the use of instrumental variables and fixed effects models can, in principle, reduce biases that would otherwise afflict a non-experimental study.

However, such natural experiments can be hard to come by. Alternatively, one might identify and control for a large number of school characteristics that might plausibly predict

program participation. This is the strategy being pursued by the Study of Instructional Improvement at the University of Michigan, which is comparing instruction and learning in three kinds of whole-school-reform efforts as well as a control condition. The Common Core of data and project data collection activities create a wealth of such school characteristics. Propensity score stratification (Rosenbaum, 1993) is a way of simultaneously controlling many potential confounders thereby plausibly reducing bias substantially.

Quasi-experimental designs capitalizing on routinely collected data. 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.

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 comparatively 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.

Increasing Precision

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. We now consider how these issues play out in a study that randomly assigns schools to alternative programs or "treatments."

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 2 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.

Insert Figure 2 Here

Figure 3 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.

Insert Figure 3 Here

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 (Puma, 1994) to estimate the variation in growth rates between and within schools for this purpose.

Enhancing precision by design. Analyses of statistical power for studies using randomization at the school level can be sobering. Given the surprisingly large number of schools required and given the typical expense of recruiting and retaining schools, study costs may be daunting. This concern naturally leads to the consideration of design options for increasing power without increasing the number of schools. One such option is the longitudinal study mentioned above. Variation between schools within will be smaller for growth rates than for student status, thus reducing the number of schools needed to achieve a given level of power. A related design is the repeated cross-sectional design. Here the object of interest is the change in school means rather than the means themselves. Changes in school means will likely vary less between schools within treatments than will the school means themselves, perhaps leading to greater power for a given number of schools. A third design option is block or match schools on variables linked to the outcome prior to randomization. For example, schools could be blocked on social and ethnic composition; randomization would proceed within blocks. This is a classic design strategy which works well when the variation between schools in the outcome is substantially smaller within blocks than when the blocks are ignored. Pre-randomization blocking can also add face validity by insuring that the treatment conditions are balanced on variables such as social and ethnic composition that are widely acknowledged to be related to outcomes.

Enhancing precision by analysis. An alternative strategy for increasing power is to identify pre-treatment covariates, at either the student or the school level, that are strongly related to the outcome. When test scores are the outcome, an obvious choice is a prior measure

of cognitive skill such as IQ or a reliable achievement pretest. Raudenbush (1997) shows that such covariates can substantially reduce the variation between schools within treatments, leading to designs with fewer schools than would otherwise be needed. Student-level covariates are particularly attractive because they result in no measurable reduction in the degrees of freedom needed for the test of significance of group differences. A downside of this strategy is that skeptics may suspect that covariates were selected post hoc to maximize the estimated treatment effect. To discourage such "cherry picking," it is essential that the choice of covariates be publicly announced prior to data collection.

Facilitating Causal Generalization

This paper has considered a conceptual model for evaluation of innovations in the field and design strategies for reducing bias and increasing precision. To the extent the suggestions made in these sections come to fruition, the result will be a stream of studies that vary in their conceptual aims and their designs. The task of "causal generalization" (Shadish and Cook, 2001) is educe as set of durable effect estimates from such a stream, where estimates of impact are hedged appropriately by knowledge of conditions that influence the effect size.

Conceptual aims. As mentioned earlier, the interpretation of an assessment of the impact of an innovation depends critically on the specific definition of the innovation and also of the comparison conditions. Cook and Campbell (1979) describe this problem as assessing the "construct validity of the causal variable." Versions of an innovation may share the same label but they may vary in important theoretical ways as may the comparison conditions. By specifying the nature of the enacted innovation and the comparison in some detail, investigators in each study enable later researchers looking across a stream of studies to assess the sensitivity of impact estimates to alternative versions of the programs being compared.

Resources. Closely related is the potentially moderating effect of resource constraints. The project of causal generalization requires an assessment of the dependence of impact estimates on resource availability (see Figure 1). This can perhaps best be done within a given study that systematically varies resources or program intensity in the fashion of a "dose-response" study in medicine. However, it will typically be necessary to cull evidence across studies that vary in resources in order to more fully assess the interplay between resources and outcomes.

Participants. The task of constructing generalizations across background characteristics of students, implementers, and settings is similar. We need to know whether the innovation works better for some kinds of kids than for others, better in some kinds of schools than others, and better when implemented by some kinds of teachers than others. To some extent, these moderating influences can be assessed in the context of a single study. For example, a study with a diverse student population can assess the interaction between the treatment and student background characteristics. And a multi-site study (see Raudenbush and Liu, 2000) can assess whether an innovation works better in some kinds of schools or districts than others. However,

once again, it will be typically necessary to synthesize evidence across studies to more fully assess these moderating influences.

Designs. As mentioned earlier (see "Reducing bias") randomized experiments can and should be done whenever causal questions are in the air and whenever such designs are feasible. However, such studies can be expensive and certain questions are less amenable than others to randomized trials. Moreover, randomized studies can degenerate into non-randomized studies as a result of attrition. As a result, a stream of research will often include studies that vary in design. It can be extremely useful to assess the dependence of impact estimates on the type of design employed. If non-randomized and randomized studies produce very similar estimates of effect, combining information across them can increase statistical power in research syntheses (c.f., Cooper and Hedges, 1994).

Power. In some cases, individual studies may lack power to detect the effects of an innovation. This will be especially true if resources for each study are limited or if study designs are expensive to implement. In this case, combining results across studies through research synthesis can substantially increase statistical power. This combining of information is most coherent when a set of studies produce results that are homogeneous. Tests of homogeneity (Hedges, 1982) are now widely used in research synthesis, otherwise known as "meta-analysis" (Glass, 1976).

Implications for Improving Evaluation Practice

I close with some final thoughts on how to improve the enterprise of systematically evaluating studies of educational innovations in field settings. The discussion above implies that a healthy scientific enterprise in education is essential if we are to be successful in developing more durable and meaningful generalizations regarding how to improve education. In part this problem of methodological infra-structure: it is essential that we increase the capacity within educational research to evaluate alternative designs, to improve data analysis, and to facilitate the synthesis of findings across streams of study. What is needed is not a simple translation of the received wisdom from biostatistics or econometrics, though these fields have a great deal to offer. Rather, educational evaluation is characterized by specific features that offer unique, interesting, and difficult methodological challenges. I have emphasized, for example, that educational innovations are essentially always implemented in varied ways by varied agents (teachers, tutors, principals, program innovators) in varied settings (tutoring sessions, classrooms, schools, and districts). This means that causal inference in education requires more flexible assumptions than is conventional in other fields. In other fields, it is routinely assumed that there is a single version of a treatment and that the treatment assignment of one participant has no impact on the potential outcomes of another participant. These notions are summarized in the "Stable Unit-Treatment Value Assumption" (SUTVA) made famous by Donald Rubin (Rubin, 1978). The agentic and social nature of educational innovation implies that SUTVA will not generally apply. Therefore alternative and more plausible assumptions are needed (Hong and Raudenbush, 2003)and these will lead to different designs than might be typically be found in

randomized clinical trials in medicine generally or pharmacology in particular.

However, enhancing methodological research and expertise in education will not be sufficient to meet the challenge discussed in these papers. These challenges arise in the interplay between educational theory, policy, and research method. Intense communication among innovators, disciplinary experts, and methodologists is essential if these challenges are to be met. This implies a need for high-quality peer review. Recent emphasis on improving peer review by foundations and by public funders such as the Institute of Education Sciences and the National Science Foundation are thus crucial as we imagine a more compelling science of innovation in education.

References

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., & Campbell, D. (1979). *Quasi-experimentation.* New York: Rand McNally.

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. D., Hunt, H. D., & Murphy, R. F. (1999). *Comer's school development program in Chicago: A theory-based evaluation.*, Northwestern University.

Cooper, H., & Hedges, L. (Editors). (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Foorman, B., Francis, D., Fletcher, J., & Lynn, A. (1996). Relation of phonological and orthographic processing to early reading: Comparing two approaches to regression-based, reading-level-match designs. *Journal of Educational Psychology, 88,* 639-652.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

Hedges, L. (1982). Fitting continuous models to effect size data. *Journal of Educational Statistics*, *7*(4), 245-270.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Hong, G., & Raudenbush, S. (2003). Causal inference for multi-level observational data with application to kindergarten retention study. Proceedings from the a*nnual conference of the American Statistical Association.* Toronto.

Puma, M. (1994). *Prospects: The Congressionally Mandated Study of Educational Growth and Opportunity.* Chicago: ABT Associates.

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

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

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

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

Shadish, W. R., Campbell, D. T., & Cook, T. D. (2001). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Co.

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

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

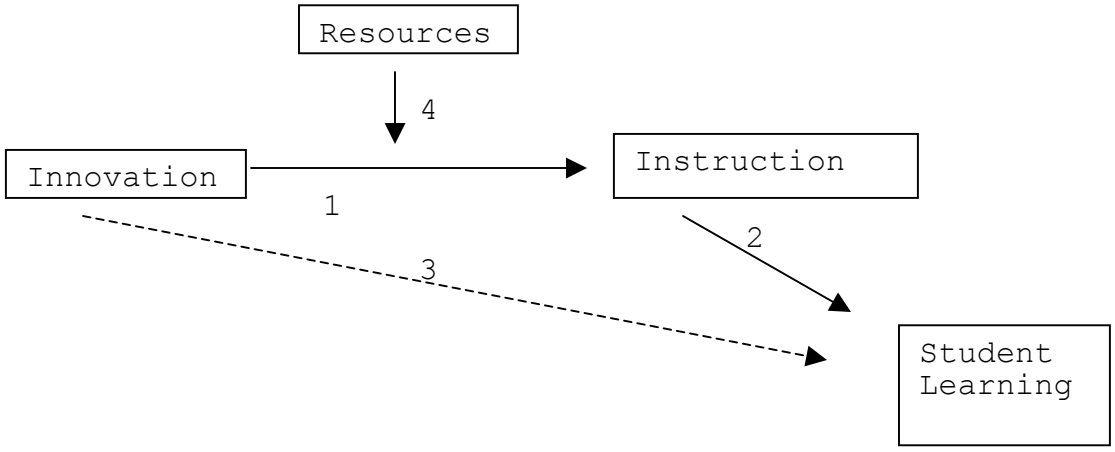

Figure 1. Conceptual Model

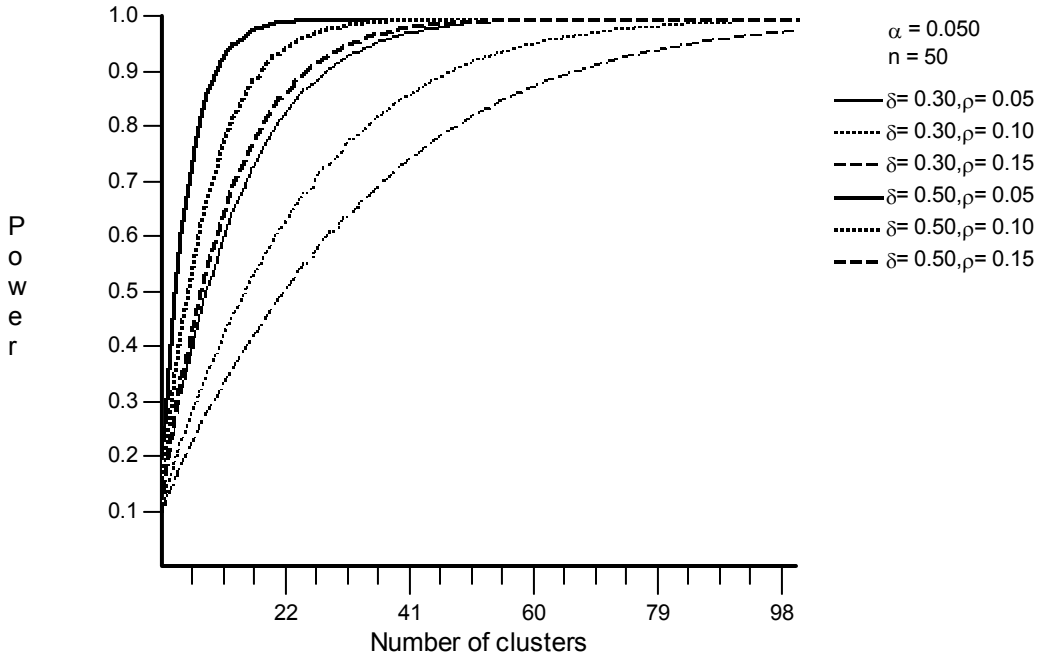

Figure 2: Power as a function of J (the number of schools) at various effect sizes and intra-school correlations), holding the within school sample size constant at n = 50.

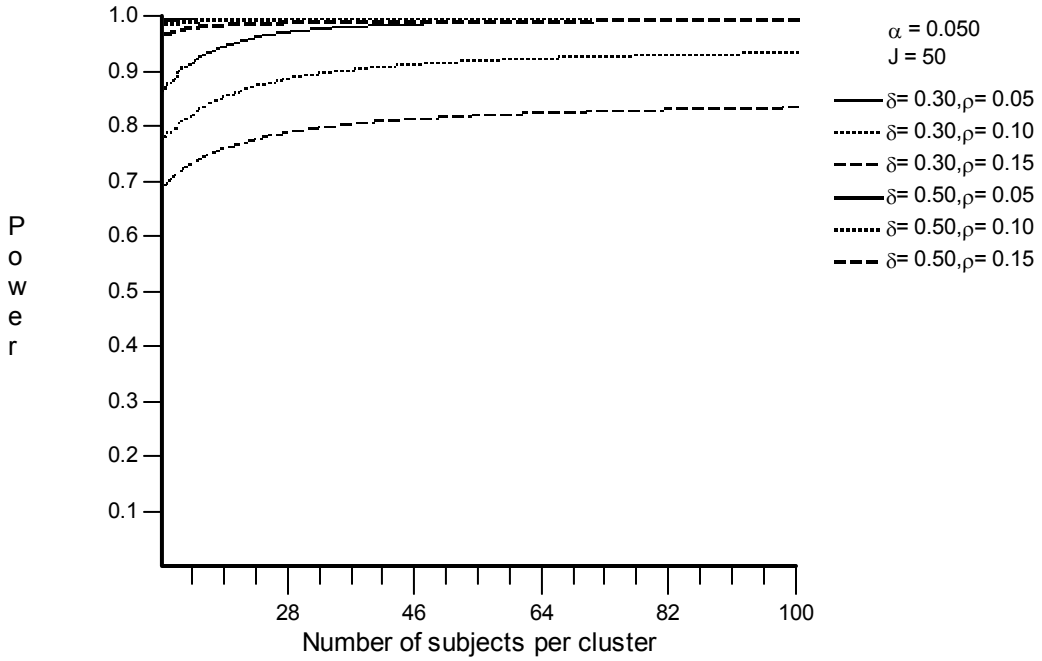

Figure 3: Power as a function of n (the number of children per school) at various effect sizes and intra-school correlations), holding the number of schools at J = 50.