

Power in Evaluations of Whole-School Change Initiatives:

Longitudinal Studies of Cluster Randomized Trials

Roderick A. Rose, MS*

Gary L. Bowen, Ph.D., ACSW

The University of North Carolina at Chapel Hill

301 Pittsboro St. CB 3550, Chapel Hill, North Carolina 27599 USA

*Tel: 1-919-962-8826, fax: 1-919-962-7557, e-mail: rarose@email.unc.edu

RUNNING HEAD: Power in Evaluations of Whole-School Change

Abstract

The calculation of power for determining sample size in a longitudinal study of a cluster randomized trial depends upon the proposed analytical model, the outcome measure, and the level at which the outcome measure is analyzed. These decisions should follow from the specification of the research question and hypothesis. We attempt to demonstrate that modest differences in analytical model design greatly influence the outcome of the power analysis and the number of clusters needed. A key implication of evaluations conducted in schools is the relatively high intracluster correlations that are found in these settings, which greatly increase the number of schools needed to satisfy statistical power requirements. Due to costs of multi-site recruitment, implementation, and evaluation such a great number of schools may not be feasible for most evaluators. We argue that there are cost-reducing options to cluster randomized trials that evaluators should consider, including variations on a quasi-experimental design.

Keywords: power analysis, multilevel, cluster randomized, longitudinal, school change.

Power in Evaluations of Whole School Change Initiatives:
Longitudinal Studies of Cluster Randomized Trials

1. Introduction

The determination of statistical power depends on assumptions about the sampling method and data structure (Cohen, 1988). However, for most evaluators, the determination of power of an evaluation is not easily accessible, even though the actual calculation of power should be simple and straightforward with the assistance of spreadsheets, statistical programs or specialized software. Little information is available and a great deal of confusion exists, regarding the decisions that must be made that ultimately determine the form of the power analysis that is conducted—particularly that power calculations are specialized to the analytical form of the evaluation model, which is generally informed by the hypotheses that will be tested (Oakes & Feldman, 2001). This implies that in multilevel modeling, the multilevel model design—for example, whether it is a two-level model or a three-level model, or whether the outcomes are subject or cluster-level—is crucial to the determination of power. However, the effect that different model designs have on the calculation of power for multilevel studies and the consequences of using the wrong model form are not clear. Raudenbush and Liu, for example, demonstrate how differences in sample design affect power for a repeated measures design, but model design itself is taken as a given (2001). The power differences demonstrated by varying the model design can be considerable. An analysis conducted with low power is a threat to statistical conclusion validity (Shadish, et al., 2002). Given this threat, a need exists for sensitivity analyses that show these effects.

These types of studies have important implications for evaluations of whole-school interventions—the focus of our work and the basis for the example used in the present discussion (Bowen, 2004). Educational intervention planners and evaluators are increasingly turning their attention to quasi-experimental and experimental research designs using assignment to treatment conditions by cluster (school) rather than subject (student), resulting in multilevel data structures (Little, 2004). Increasing emphasis is also being placed on evaluation in the context of development—using multiple time points to assess change rather than simple pre/post designs (Willet, et al., 1998). A study of the statistical power of the design, to detect the program effect prior to conducting the evaluation, is necessary to ensure that enough schools are sampled (Bloom, et al., 1999).

This study uses a sensitivity analysis to demonstrate how minor changes in the model design affect the calculation of power, the outcome of the power calculation, and ultimately the decision about how many schools to sample. After a brief review of the issue of power, we review the methods and requirements for using a power calculation to determine the cluster sample size in the context of multilevel evaluation designs, including how to prepare to conduct a power analysis. Next, we conduct a sensitivity analysis comparing the power calculation for three multilevel models with different model designs and present the results of this analysis. We conclude by discussing the advantages and disadvantages of the three models and by establishing a context for further investigation.

Generally, the methods we propose have applications to many different types of clustered settings—e.g., communities or families. The implications discussed in the conclusion are more specialized for whole-school settings due to the great expense of

sampling schools but may be applicable to some other settings, such as communities, with similarly large clusters.

2. The Issue of Power

Power is affected by decisions about the method of analysis, the analytical model, the proposed outcome measure, and the level at which that outcome measure is analyzed. These decisions—which become assumptions or variables in the power analysis—are strongly related both to the theory of change of the program under study and to the associated research questions and hypotheses (Oakes & Feldman, 2001; Cohen, 1998; Maxwell, 1998; Muller, et al., 1992). A single set of data may give rise to different assessments of power, depending upon the evaluator's decisions.

Generally, the structure of the data informs the selection of analytical method, and a three-level heuristic of time-subjects-clusters has been developed for modeling longitudinal studies of cluster randomized trials (Bryk & Raudenbush, 1988). However, there also are appropriate two-level models. For example, a model of school aggregates over time may answer questions relevant to school administrators whose schools must meet standards that are not wholly dependent upon the progress of individual children.

Two-level alternatives have great appeal because they appear to be simpler or more practical, yet this appeal may be misleading. The simpler formulations of these types of models may lead evaluators to believe that, everything else being equal, fewer schools will be needed to test the hypothesis, but this is not necessarily the case. Thus, if power is dependent on the assumption of model design, then calculating power for one of these simpler formulations may result in the wrong number of schools being sampled, which

may be as bad as conducting no power analysis at all (Muller et al., 1992). A sensitivity analysis of a set of alternative specifications reveals unexpected limitations of some designs and, when combined with theoretical and practical needs of the evaluation, demonstrates their relative value to understanding a program's impact. We now review how the calculation of power is affected by the multilevel nature of the evaluation of a longitudinal study of a cluster randomized trial.

3. Power and the Evaluation of Multilevel Intervention Programs

Both cluster randomization and longitudinal design elements result in multilevel data, which should therefore not be planned or analyzed using methods appropriate for single-level data. In cluster randomized designs, the data are multilevel because the unit of randomization and the unit of analysis are not at the same level (Donner, 1998). This introduces complex error structures that multivariate normal models cannot handle. Cluster randomization introduces variability in cluster-level characteristics—commonalities shared by members of a cluster (who are not selected randomly into their clusters) that can be generalized to the cluster level (Aitken & Longford, 1986). Cluster randomization is appropriate for evaluations of whole-school programs because treatment and control subjects from the same cluster would lead to contamination.

Longitudinal data are multilevel because they are represented by time-varying observations “within” subjects, with time-invariant characteristics measured at the subject level; these also have complex errors (Willett et al., 1998). Longitudinal and cluster randomized multilevel data structures should be analyzed using appropriate methods, such as hierarchical linear modeling (HLM) (Bryk & Raudenbush, 1988) or latent variable

analysis (Muthén, 2004), which allow for these complex error structures. As a consequence of combining the two design elements—cluster randomization with longitudinal trajectories—the data are characterized by three levels: time, subjects, and clusters (the overlapping level being the subject). We now turn to discuss the calculation of power in multilevel settings.

3.1. Power in Multilevel Settings

The study of power for a program effect in multilevel settings concerns variance, sample size, significance level (one-tailed), and effect size, just as in a conventional setting (Cohen, 1988). The multilevel framework simply multiplies the number of variances and sample sizes needed to inform the power of a given effect by the number of levels in the design (Bloom et al., 1999; Raudenbush, 1997; Raudenbush & Liu, 2001). As we noted earlier, we are primarily interested in the number of clusters to be sampled. A minimum at this level is desirable because a high cost is often associated with sampling additional clusters, particularly in school and community settings.

The concept of a “minimum” indicates optimality, that the number of clusters sampled results in minimum standard error for the program effect, given budgetary constraints and the assumptions employed in the design (Snijders & Bosker, 1993). Determining this minimum is the goal of a power analysis, which we refer to in shorthand as calculating “power for sample size”. For the calculation of power to go forward, variances and effect size must be estimated—perhaps by using pilot data or a review of literature, which may not always be available (Donner, 1998). In such cases, an informed guess, with a few weak assumptions, may be needed (Muller et al., 1992).

3.2. Calculation of Power

The calculation of power for cluster randomized designs can be conducted using the guidelines set forth in Bloom (1995) and Raudenbush (1997). For longitudinal or repeated measures designs, the guidelines found in Raudenbush & Liu (2001) can be used. We have developed a SAS program library that calculates power for linear contrasts, using these guidelines. For most evaluators, power will be calculated using one of many statistical programs, such as SAS. There are many free power calculators available but not all of them will calculate power for three-level models. A free graphics-based program called Optimal Design (developed by Stephen W. Raudenbush, Xiao-Feng Liu, and Richard Congdon), which will calculate power for longitudinal cluster randomized trials, is available from Scientific Software International (<http://www.ssicentral.com>). The SAS library and Optimal Design produce effectively equal results. The SAS library is available on the “Analytical Tools” section of the School Success Profile web site, <http://www.schoolsuccessprofile.org>.

4. Planning a Power Analysis

Conducting the analysis may be fairly straightforward with the help of statistical programs once the inputs to the power calculation are determined. But determining these inputs represents a critical step—the assumptions and design elements informing the analysis plan for an evaluation have a substantial effect on the outcome of the power analysis and may result in different optimal samples sizes. It is important to prepare in advance how the data will be analyzed, while the study is forming and long before any data are available. The objective of this preparation is to identify and parameterize an impact

measure as a testable model parameter, an effect estimate representing the difference in the outcome between experimental and control or comparison groups (Snijders & Bosker, 1999). Identifying this impact measure requires deciding on an outcome against which to judge this program, which will further help determine the structure of the analytical model and an appropriate method of analysis; these must also be known during the planning stages of the power analysis (Cohen, 1998; Maxwell, 1998).

Finally, all of these decisions should make sense in the context of the theory of change or the logic model of the program. The theory of change or logic model indicates how the specified actions to be taken affect the intermediate and ultimate outcomes of the intervention program (Julian, 1997). A logical first step in planning the power analysis, therefore, is to specify the research questions and hypotheses to be tested. These questions and hypotheses should address the causal pathway specified in the theory of change, from contexts to causes to intermediate and ultimate outcomes. Even as they pertain to the ultimate outcome, several research questions and hypotheses may be available; this is particularly true in multilevel settings in which a single measurement instrument can be measured at different levels through aggregation or disaggregation (Hox, 1995).

4.1. Aggregation and Outcome Level

If the theory of change specifies that the program has an effect on subjects, then the ultimate outcome measured and tested in the model should be at the subject level.

Alternatively, if the theory of change specifies that aggregative measures are affected but does not specify individual subject effects, then the outcome should be measured and tested at the cluster level. The two measurements are different, even if they are based on

the same instrument. Changes in end-of-grade test scores at the student level measure developmental trajectories of individual youths. Changes in end-of-grade test score composites at the school level, on the other hand, may measure a combination of developmental trajectories and cohort replacement or attrition (Gail et al., 1996).

4.2. Misalignment

Failure to analyze power in the context of the model used for the data analysis may result in misalignment—the “right answer to the wrong question” (Muller et al., 1992, p. 1209). The authors note that in these situations the power analysis does not adequately reflect conditions that are modeled, which may result either in an insufficient sample size or in costly over-sampling (which is not optimal). This applies to the sampling design as well as to the appropriateness of the research question for assessing the effectiveness of the program, as demonstrated in the theory of change.

In this study, we are attempting to demonstrate how slight variations in these design aspects and a failure to adhere to the recommendations can lead to different cluster sample sizes and misalignment of the power analysis with the evaluation it is intended to represent. A sensitivity analysis is used to compare a set of different specifications.

5. Method

We are interested in calculating power to determine the number of schools needed to evaluate a whole-school intervention program. For this analysis, we will consider the calculation of power for an existing program that is currently being evaluated in a quasi-experimental setting in 11 North Carolina schools. Recall that the theory of change is important to preparing the power calculation; in this case, the program attempts to change

school organizational culture by targeting the staff, faculty, administrators and other stakeholders, in order to ultimately promote greater student achievement (Bowen, 2004). This is whole-school, making it impossible to have within-school control subjects, thereby requiring us to sample by cluster.

For the sensitivity analysis, we contrasted three different models while holding the sampling method, data structure, and method of analysis (HLM) constant. We used student end-of-grade exams as the measurement instrument. Scores on the end-of-grade exams are developmental scale scores, allowing us to use them as repeated measures. The three models tested variations in the dependent variable between individual student and school composites, and variations between the longitudinal cluster-randomized model and a simplified two-level version of individual growth over time (ignoring the clustering). Because the outcome level varied between student and school and was therefore measured on a different metric, we used a standardized effect size of .5 for comparison purposes. We used the method that was appropriate to the calculation of power for the specified structure of the model, which differs slightly across the models. We took the slightly different methods of calculating power for the two- and three-level models as a given.

We used a consistent notation throughout all of the models. The subscript j was used to indicate clusters, i for subjects, and t for the four time points (from zero to three). Because end-of-grade math scores tended to fluctuate more from year to year, and therefore had higher standard deviations, we chose to report the results using end-of-grade math scores rather than end-of-grade reading scores (which understated the number of schools that would be needed to analyze math scores). We used unconditional linear

growth models (in which time was the only regressor). No program effects were entered into the models. All models were estimated using SAS Proc Mixed.

5.1. Model 1: School-level Trajectories

School-level trajectories were constructed using serial cross sections of student achievement on end-of-grade math exams. Taking the mean across all students with an exam score created a longitudinal trajectory of school outcomes. These trajectories represented overall school performance over several years and were an inseparable mix of student developmental trajectories, cohort replacement, and attrition (Gail et al., 1996). The program impact (D_1) was estimated as the difference in average trajectories between schools receiving and those not receiving the program. The program effect for this model represented the difference in the average growth of these annual aggregates between each condition. The hypothesis tested was that the program has a positive impact on overall school performance, when compared to schools not receiving the program.

Model 1 was a two-level model. Level 1 took the following form, in which $SCORE_{tj}$ is the mean end-of-grade math score for school j in period t :

$$(1) \quad SCORE_{tj} = \beta_{0j} + \beta_{1j} TIME_{tj} + r_{tj}$$

$$Mean(r_{tj}) = E(r_{tj}) = 0; \text{ Variance}(r_{tj}) = V(r_{tj}) = \sigma^{*2}$$

Level 2 (school) took the following form, modeling initial condition and growth on school-level means and random components:

$$(1.0) \quad \beta_{0j} = \pi_{00} + u_{0j}$$

$$(1.1) \quad \beta_{1j} = \pi_{10} + u_{1j}$$

$$E(u_{0j}) = 0; V(u_{0j}) = L; E(u_{1j}) = 0; V(u_{1j}) = M$$

$$\text{Covariance } (u_{0j}, u_{1j}) = \text{Cov } (u_{0j}, u_{1j}) = \psi$$

5.2. Model 2: Individual Growth (Two-level)

This model used measures at the student level over time, not aggregated to any higher unit level. However, the clustering of students within schools was ignored in the model. The model examined longitudinal-developmental trajectories of students based on achievement on end-of-grade math exams. We estimated the program impact (D_2) as the difference in average trajectories between students receiving and those not receiving the program (disregarding school membership, though it is implicit that students “receiving the program” were actually *enrolled in schools* receiving the program). The program effect represented the difference in average student growth between each treatment condition. The hypothesis tested was that the program has a positive impact on student achievement, when compared to the achievement of students not receiving the program.

The Level 1 (time-level) model took the following form, in which the dependent variable SCORE_{ti} is the end-of-grade math score for student i in period t :

$$(2) \quad \text{SCORE}_{ti} = \beta_{0i} + \beta_{1i} \text{ TIME}_{ti} + r_{ti}$$

$$E(r_{ti}) = 0; V(r_{ti}) = \sigma^2$$

The Level 2 (student-level) model took the following form, regressing each time-level parameter on student means and random coefficients:

$$(2.0) \quad \beta_{0i} = \pi_{00} + u_{0i}$$

$$(2.1) \quad \beta_{1i} = \pi_{10} + u_{1i}$$

$$E(u_{0i}) = 0; V(u_{0i}) = W; E(u_{1i}) = 0; V(u_{1i}) = T$$

$$\text{Cov } (u_{0i}, u_{1i}) = \zeta$$

5.3. Model 3: Individual Growth within the School Context (Three-level).

Model 3 began (much like Model 2) with longitudinal-developmental trajectories of students based on end-of-grade math exam achievement. However, in this case the clustering of students within schools was not ignored. The student-level model parameters were modeled on school-level characteristics and random effects, and a between-school program impact (D_3) was estimated as the difference in average trajectories of students in schools with or without the program.

A useful heuristic for understanding what the program effect represents in this model is to consider that it was measured by first averaging the within-school individual growth rates for each school, then subsequently averaging these parameters across schools within each treatment condition and taking the difference between conditions. This heuristic can be contrasted to the program effect measured in Model 1, in which individual scores (rather than growth rates) were averaged and then trajectories of these averages were modeled across schools for each treatment condition. The hypothesis tested in Model 3 was that the program has a positive impact on student achievement in schools receiving the program, when compared to student achievement in schools not receiving the program.

The Level 1 (time) model took the following form, in which $SCORE_{tij}$ is the math score of student i (who is enrolled in school j) in period t :

$$(3) \quad SCORE_{tij} = \beta_{0ij} + \beta_{1ij} TIME_{tij} + r_{tij}$$

$$E(r_{tij}) = 0; V(r_{tij}) = \sigma^2$$

The Level 2 (student) model regressed time-level parameters on individual student means and random coefficients:

$$(3.0) \quad \beta_{0ij} = \pi_{00j} + u_{0ij}$$

$$(3.1) \quad \beta_{1ij} = \pi_{10j} + u_{1ij}$$

$$E(u_{0ij}) = 0; V(u_{0ij}) = S; E(u_{1ij}) = 0; V(u_{1ij}) = T$$

$$\text{Cov}(u_{0ij}, u_{1ij}) = \zeta$$

The Level 3 (school-level) regressed the individual means on school means and random coefficients:

$$(3.0.0) \quad \pi_{00j} = \gamma_{000} + e_{00j}$$

$$(3.1.0) \quad \pi_{10j} = \gamma_{100} + e_{10j}$$

$$E(e_{00j}) = 0; V(e_{00j}) = L; E(e_{10j}) = 0; V(e_{10j}) = M$$

$$\text{Cov}(e_{00j}, e_{10j}) = \psi$$

5.4. Estimating Inputs to Power

In order to estimate the power for sample size for these three models, the variances in the growth parameters from each model (T and M) and the error variance (σ^{*2}), which were all unknown, had to be estimated. The variances in the intercept parameters, as well as the covariances, could be safely ignored in the calculation of power because we were interested only in the variance in growth. Because the estimates from the actual program were not available, another source had to be used. We explored two options that were recommended by the literature on power analysis—literature review and calibration study. A third option, using pilot study data (Goldstein, 1987), was not available for this program; although a pilot study was conducted, no data were available at the time.

5.4.1. Literature review

It is ideal that the values of the variance components be based on the results of previous research (Bloom, 1995; Maxwell, 1998). We searched for variance and sample size estimates obtained during the study of other programs, and in the absence of such data, we examined general studies of organizational culture on academic outcomes. Similar models are unfortunately rare in literature, and it is even rarer to report the variance estimates that are needed in this situation.

5.4.2. Calibration study

A calibration or simulation study can be conducted if data on a similarly styled whole-school intervention program are available (Goldstein, 1987). The strategy is to estimate the models using these data and to use the variance components as inputs to power for sample size. We chose to use this option because data on a similar intervention program were available in an administrative database of academic outcomes, provided by North Carolina that would be used for the study. The state's program is similar to our program in that it targets organization-wide change.

The state of North Carolina administered this intervention program to schools that demonstrated low performance over a certain time frame. Using the data from this program, which was administered to 14 schools in 1998, and adding to this group a balanced sample of similar low-performing schools that did not receive the program in the same year, we were able to analyze all three models, using hierarchical linear modeling. For computational ease, the covariance terms at the student and/or school levels were

constrained to zero. (This was required because Proc Mixed was not able to estimate a three-level model with a free covariance parameter given these data.) We then inserted the pooled variances from this model into the power calculation for each model. The low number of schools in each treatment condition probably inflated the variances, making the power estimates conservative, which is preferred.

5.5. Power Analysis

After obtaining the needed variance components, we conducted the power calculation, combining these variances with other inputs. Several inputs were fixed across all three models. We assumed the inputs for Level 1 sample size (time; 4 time points); for Level 2 sample size (200 subjects per school); for the level of significance (one-tailed $\alpha = .05^1$); and for power ($1-\beta = .80$). A standard deviation effect size of .5 was used, based on Cohen (1988). The analytical method (multilevel modeling or HLM) and the theory of change, which was based on a program model of whole-school organizational change influencing individual student achievement, were also constant across all three models. In our estimate of these power calculations, we also chose to assume a balanced design for the treatment conditions.

For Models 1 and 3, we used the power analysis to directly estimate the number of schools to sample. In the case of Model 2 (individual growth over time), we estimated the number of students needed and then derived the number of schools on the basis of our assumption of 200 subjects per school. Models 1 and 2 use a strictly longitudinal design and were estimated using the SAS library and confirmed with the Optimal Design

¹ For Optimal Design, the α represents a two-tailed test; we entered $\alpha = .10$, based on the approximation of a one-tailed .05 to a two-tailed .10.

software. Model 3 uses an individual growth-within-clusters approach and was estimated using the same tools.

6. Results

6.1. Calibration Study

The results obtained by using the chosen calibration data to estimate variances for each of the three models appear in Table 1. The table shows each of the three models on the rows; each of three variance components is reported on the columns. Between-student variance in growth is not applicable for Model 1 or Model 2, and the metric for this measurement differs between Models 1 and 3. In Model 1, between-school variance represents the variation between schools in a trajectory of school averages; in Model 3, it represents the variation between schools in average student trajectories (i.e., a trajectory of averages in Model 1, versus an average of trajectories in Model 3).

The difference in magnitude of the error and parameter variances between Model 1 and Models 2 and 3—with the Model 1 estimates being much smaller—illustrates the loss of information from pre-aggregating the data before modeling trajectories. Note also the similarities between Models 2 and 3, in which a portion of residual variance in Model 2 (equal to 1.857) appears as between-school variance in Model 3. This was the only outcome difference between the two models, and this had a substantial effect on the calculation of power. The appropriate variances were inserted into the power calculation formulas for the linear contrast repeated measures design (Models 1 and 2) and the individual growth-within-clusters design (Model 3).

[Insert Table 1 about here.]

6.2. Power Analysis

The results of the power analysis appear in Table 2. The Model 1 design indicated that a minimum of 125 schools would need to be sampled, given our assumptions. The Model 2 design indicated that 537 students would be required, which is a minimum of three schools. However, because we have assumed 200 students per school and have proposed a balanced design with an equal number of schools in each experimental condition, this would require four schools—two in the experimental group and two in the control group. Finally, the Model 3 design demonstrated that a minimum of 68 schools would be required. These findings represent a balanced approach—half the sample in the experimental group and half in the control group.

[Insert Table 2 about here.]

7. Conclusions and Discussion

The outcome of a power analysis for a program effect in a multilevel study depends on a series of evaluator decisions, assumptions, and guesses. This includes decisions about model design. In a longitudinal study of a whole-school intervention program, the most appropriate model design may be a three-level model with time at the first level, students at the second level, and schools at the third level (Bryk & Raudenbush, 1988). Other models, such as a two-level model of school-level aggregates measured over time (with time at the first level and school characteristics at the second level), may be appropriate in this setting, depending on the research question and the hypotheses under study. However, though they are simpler in design, these other models will not necessarily help to reduce the sample

size needed to achieve the desired level of power for testing the program effect, and they should not be thought of as a strategy toward achieving that goal.

Findings from the present analysis lend support to this conclusion. Each proposed model provides a very different answer to the question of power for sample size. All else being equal, the model selected should be the one most appropriate to the parameters of the program and the effect we want to demonstrate, and the model which is sufficiently rigorous, in addition to practical considerations. For example, Models 1 and 2 were both simpler, two-level models for measuring the effects of a whole-school intervention program. Though from a strictly economic point of view it would be ideal to choose Model 2, because only four schools would appear to be required, one of the model's disadvantages is that it treats all students as independent, which was not a characteristic of the data. Assuming independence and ignoring the clustering of students within schools discarded without consideration a major source of variation between students, and therefore variance was attributed to the wrong source. In fact, as demonstrated by model 3 (which is otherwise identical but captures this effect), 64% of variance is at the school level.

Indeed, this result has previously been observed in school settings. In a study of 618 students in 86 schools, Bryk & Raudenbush, (1988) found that nearly 83% of the variation in students' growth in math achievement was between schools. In the present study, the reduced standard errors and inflated degrees of freedom for the t tests would yield spurious findings of significance. Effectively, with a sample size of only four schools, there would be insufficient evidence to attribute observed differences to the

program, even though the statistics might indicate it, because the differences could be attributable to the school.

Trajectories of entire schools over a multiyear period, regardless of cohort retention or attrition, are an important indicator of school performance, and these trajectories were modeled in Model 1. However, Model 1 proposes an impractical sample size of 125 schools. The number of time points could be greater because we would not be limited to individual students' time in each of the schools—we could use up to 10 years' data for model 1. A further analysis demonstrated that increasing the number of time points reduces this minimum sample size, but not satisfactorily (i.e., not to a level comparable with Models 2 or 3). Therefore it is not a recommended “second best” solution for demonstrating the program effect—it does not improve the power at a cost advantage. In addition, though Model 1 does test a question that may be of relevance, it does not test the hypothesis of an effect on student outcomes—an important part of the program model and of our evaluation needs. The aggregation of all student data to the school level and the analysis of the program at this level made it impossible to estimate within-school differences.

Model 3 requires sampling a minimum of 68 schools. Using this model, it would be possible to study the program effect both on schools and on individual students. It did not inflate the degrees of freedom or depress standard errors (like Model 2); it also did not throw away important variability and lose precision (like Model 1).

Model 3 does have disadvantages and limitations. First, a three-level model of individual change within clusters is limited to one type of clustering (e.g., one cannot study

school effects and neighborhood effects at the same time). Added levels and cross-clustered models are theoretically possible (Raudenbush & Bryk, 2002), but though the calculations for such models may eventually become possible as computing power increases to the needed threshold, it is not yet practical. Finally, 68 schools may be a prohibitively large and expensive sample to recruit, treat and study. The high sample size is strongly related to the high proportion of variance attributable to the school level (the intraclass correlation)—according to the calibration model, this is 64%. This has led us to propose an alternative quasi-experimental study, which we discuss in section 8.

In conclusion, using only a very modest difference in analytical design, this study demonstrates that evaluators must think through their entire set of assumptions and decisions when attempting to identify a concise measure of power for sample size, a process which requires advanced knowledge of analytical design and evaluation methods and models. Evaluators must be acutely aware of the consequences of seemingly minor variations in design and analysis planning. Alignment of the power analysis with the eventual analytical design for the study, which can only be accomplished through advanced planning of the analytical design, is crucial to avoiding over-sampling or failing to sample enough clusters.

8. Implications for Evaluations of Whole-School Programs

The intraclass correlation in the calibration study used to inform the power analysis was .64, which is large, leading to a minimum sample of 68 schools. In the method of calculating power for a longitudinal cluster randomized trial (model 3 in this study), the intraclass correlation (ICC) dominates all other inputs. For example, if the

ICC were only .20 (as estimated in a misinformed earlier version of this analysis), we would need to sample only 22 schools. Similar variations (by a factor of about 3) in other inputs have modest and sometimes imperceptible effects on the optimal sample size. Furthermore, this is given a “medium” effect size, which is a more liberal estimate than many evaluators might be comfortable with in whole-school situations: it supposes that the average periodic (annual) effect of the program will be a half-standard deviation. However, a more-realistic standardized effect of one-fifth of a standard deviation would require us to sample an unrealistic number of more than 400 schools. It is well-documented that cluster randomized trials, for this reason alone, have very low power to detect effects (Little, 2004).

There are many documented strategies that can be employed for enhancing the power of a given design (Shadish, et al., 2002). One method is to use an unbalanced sample, which would be possible in settings in which there is a cost difference between sampling treatment schools and sampling non-treatment schools (Gail et al., 1996). Unbalanced samples have lower power than balanced samples of equal total sample size. However, if there is an economy to increasing the comparison sample, and if the treatment schools represent the primary cost constraint, the total sample can be increased by sampling more comparison schools and reducing the treatment sample. Other strategies that can be used to increase power or decrease the needed sample size include the matching of experimental to control units (Donner, 1998), and the analysis of conditional models that control for variation by using covariates to model fixed effects (Maxwell, 1998; Raudenbush, 1997).

As noted earlier, the goal of a power analysis is optimization—specifically, sampling a sufficient, minimum, number of schools given a cost constraint. In this study, we have primarily considered the sample size issue but have ignored the cost issue. There are legitimate questions about the cost of conducting an evaluation, and strategies that can reduce cost. The expense of conducting a cluster randomized trial in schools includes the cost of the treatment as administered, the conduct of an evaluation across multiple sites (in both treatment and non-treatment sites) and the recruitment of both treatment and non-treatment school sites. When we recently conducted an evaluation of a whole-school change initiative, we selected our sites for recruitment on the basis of district; this reflects the political realities of attempting to recruit schools. We began with four districts containing an insufficient number of eligible schools for both the treatment and comparison samples—there were only enough for treatment. We did this because the comparison sample data could be obtained from the database provided by the state of North Carolina at zero marginal cost, and there would be no need to establish contact with them. We therefore had to recruit only treatment sites. This made our evaluation plan quasi-experimental because schools would not be randomly assigned to treatment conditions.

However, by limiting our recruitment to treatment schools only, we addressed two concerns: we reduced the direct cost of recruitment of the needed number of schools because we did not have to travel to or prepare recruitment materials for non-treatment schools, which—according to our unbalanced power estimate—outnumbered our treatment schools three-to-one. Also, by guaranteeing to all of the sites that we contacted that they

would be included in the treatment group if they consented to participation, we also reduced the pool of recruits that we would have needed to ensure a sufficient number of consenting districts and schools. Indeed, even given the 100% guarantee of treatment in our evaluation, only 50% of the districts consented to participate. We find it unlikely that as many districts would have consented if their schools had been promised only 50% certainty of membership in the treatment sample—it was a question raised by district personnel in both districts consenting to participate. One strategy that we considered in order to encourage higher participation in this situation was to conduct a staggered control—introduce the treatment to the control schools after a period of time (Little, 2004). However, in order to observe a program effect, the treatment lag would have to be as long as the amount of time needed to sufficiently diffuse the program throughout the school (and thus observe the program effect), which may be three years or longer, an unappealing length of time to wait to receive the treatment.

Finally, the number of districts and sites required and their geographic heterogeneity would have been prohibitively costly beyond the recruitment stage. Consequently, we opted for the quasi-experimental design, using two districts neighboring each other—one rural and one urban. The quasi-experimental design did not increase power, and it also introduced the possibility of additional challenges to internal validity (Shadish, et al., 2002), many of which may have plausibly been operating (this has not been tested or studied in detail yet). However, it relaxed the cost constraints on power, thus making it possible to sample the number of schools required by the power analysis.

Acknowledgements

This research is based on data from the North Carolina Education Research Data Center at Duke University, directed by Elizabeth Glennie and supported by the Spencer Foundation. The authors also wish to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information.

References

- Aitkin, Murray and Nick Longford. 1986. "Statistical Modelling Issues in School Effectiveness Studies." *Journal of the Royal Statistical Society Series A (General)* 149(1):1-43.
- Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5):547-556.
- Bloom, Howard S., Johannes M. Bos, and Suk-Won Lee. 1999. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs." *Evaluation Review* 23(4):445-469.
- Bowen, Gary L. 2004. "Quality Youth Interventions through Community Assessment: Evaluations of the School Success Profile Intervention Package." *The Evaluation Exchange* 10(1):27.
- Bryk, Anthony S. and Stephen W. Raudenbush. 1988. "Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model." *American Journal of Education* 97:65-108.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). Hillsdale, NJ: Erlbaum.
- Cohen, Michael P. 1998. "Determining Sample Sizes for Surveys With Data Analyzed by Hierarchical Linear Models." *Journal of Official Statistics* 14(3):267-275.
- Donner, Allan. 1998. "Some Aspects of the Design and Analysis of Cluster Randomized Trials." *Applied Statistics* 47(1):95-113.

- Gail, Mitchell H., Steven D. Mark, Raymond J. Carroll, Sylvan B. Green, and David Pee. 1996. "On Design Considerations and Randomization-Based Inference for Community Intervention Trials." *Statistics in Medicine* 15:1069-1092.
- Goldstein, Harvey. 1987. *Multilevel Models in Educational and Social Research*. London: Oxford University Press.
- Hox, J. J. (1995) *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.
- Julian, David A. 1997. "The Utilization of the Logic Model as a System Level Planning and Evaluation Device." *Evaluation and Program Planning* 20(3):251-257.
- Little, Priscilla M. D. 2004. "A Conversation with Howard Bloom and Stephen Raudenbush." *Harvard Family Research Project: The Evaluation Exchange* 10(1):16-17.
- Maxwell, Scott E. 1998. "Longitudinal Designs in Randomized Group Comparisons: When Will Intermediate Observations Increase Statistical Power?" *Psychological Methods* 3(3):275-290.
- Muller, Keith E., Lisa M. LaVange, Sharon L. Ramey, and Craig T. Ramey. 1992. "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications." *Journal of the American Statistical Association* 87(420):1209-1226.
- Muthén, Bengt. 2004. "Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data." Pp. 345-368 in *Handbook of Quantitative Methodology for the Social Science*, edited by D. Kaplan. Newbury Park, CA: Sage Publications.

- Oakes, J. M. and Henry A. Feldman 2001. "Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models." *Evaluation Review* 25: 3-28.
- Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods* 2(2):173-185.
- Raudenbush, Stephen W. and Xiao-Feng Liu. 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.
- Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Snijders, Thomas A. B. and Roel J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
- Snidjers, Thomas A. B. and Roel J. Bosker. 1993. "Standard Errors and Sample Sizes for Two-Level Research." *Journal of Educational Statistics* 18(3):237-259.
- Willett, John B., Judith D. Singer, and Nina C. Martin. 1998. "The Design and Analysis of Longitudinal Studies of Development and Psychopathology in Context: Statistical Models and Methodological Recommendations." *Development and Psychopathology* 10:395-426.

Table 1. Results from Calibration Study (Inputs to Power Obtained from Models 1, 2, and 3).

<u>Model</u>	Residual variance (σ^2)	Between-student variance in growth (T)	Between-school variance in growth (M)
(1) School two-level	0.591	N/A	0.464
(2) Student two-level	23.510	1.052	N/A
(3) Student three-level	21.720	1.052	1.857

Table 2. Results from Power Calculation (Optimal Sample Sizes).

<u>Model</u>	<u>N (Number of students)</u>	<u>J (Number of schools)</u>
(1) School two-level	200 per school ^{&}	125
(2) Student two-level	537	4*
(3) Student three-level	200 per school ^{&}	68

*Assuming 200 students per school and a balanced design, $N = 537$ implies that four schools should be sampled.

[&]Assumed.