Title: Improving the Design of Science Intervention Studies: An Empirical Investigation of Design Parameters for Planning Group Randomized Trials

Authors and Affiliations:

Carl Westine
Western Michigan University

Jessaca Spybrook
Western Michigan University

Abstract Body

Background:

In the past decade, there has been a dramatic increase in the number of group randomized trials (GRTs) designed to test the effectiveness of educational programs, policies, and practices (Authors, 2012). In order for these GRTs to yield high-quality evidence of whether or not a program is effective, among other things, the study must be well designed with adequate power to detect a treatment effect of a reasonable magnitude. The field has made substantial progress in terms of how to calculate statistical power for GRTs (e.g. Donner & Klar, 2000; Konstantopoulos, 2008, Raudenbush, 1997; Raudenbush & Liu, 2001; Authors, 2007; Schochet, 2008). In addition, there has been new empirical work with respect to the design parameters necessary for statistical power calculations for GRTs for reading outcomes, math outcomes, and cognitive outcomes for preschoolers (Bloom, Richburg-Hayes, & Black, 2007, Hedges & Hedberg, 2007; Jacob, Zhu, & Bloom, 2010; Schochet, 2008). However, there is little empirical information available for design parameters for science outcomes. Further, it is unclear whether or not the empirical estimates for reading and math design parameters transfer to science for reasons such as the fact that unlike reading and math, science is not typically tested annually. GRTs of science interventions are becoming more common and hence it is critical to build a base of empirical design parameters specific to science outcomes.

Purpose:

The purpose of this study is to calculate empirical estimates of design parameters for science outcomes to improve the design of GRTs of science interventions. Using statewide data from Texas, we seek to:

1) calculate the unconditional intraclass correlations (ICCs) for science outcomes for three types of models including the:
   a. 2-level hierarchical linear model (HLM) with students nested within schools,
   b. 3-level HLM with students nested within schools nested within districts, and
   c. the within district 2-level HLM with students nested within schools.
2) calculate the percent of variance explained (R-squares) when the following sets of covariates are included in the models:
   a. student-level demographics,
   b. student-level pretest,
   c. school-level pretest, and
   d. student-level pretest and school-level pretest.

Data:

The data for this study comes from the state of Texas. In accordance with FERPA policies, the Texas Department of Education provided masked K-12 student data for five academic years. The key variables of interest included in the dataset are student ID, race, gender, disadvantaged status, science scores, math scores, reading scores, school ID, and district ID. Science is tested in grades 5, 8, 10, and 11 in Texas. Reading and math are tested annually.

The dataset includes more than 900 districts with approximately 3,500 elementary schools, 1,600 middle schools, and 1,300 high schools. The number of student records available for each grade in each year ranged from 273,000 to 373,000, with an average of 326,000. Due to masking and missing data (i.e., alternative test format, lack of complete demographic information), sample sizes for each grade in each year were reduced by approximately 30 percent, with masking and missing data each contributing about half of the total data reduction.

Statistical Models and Analysis:

There are three primary models of interest for this study. The first is the 2-level hierarchical linear model (HLM) with students nested within schools. This model ignores the district level. The unconditional models are as follows (Raudenbush & Bryk, 2002).

The level-1, or student-level model is:

$$Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2) \qquad (1)$$

where y_{ij} is the outcome for individual $i = \{1,..., n\}$ in school $j = \{1,..., J\}$; β_{0j} is the average at school *j*; and e_{ij} is the residual error associated with each student.

The level-2 model, or school-level model is:

$$\beta_{0j} = \gamma_{00} + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00}) \qquad (2)$$

where γ_{00} is the grand mean and τ_{00} is the variance between schools. In this case, the ICC is the between school variance relative to the total variance, or $\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$.

The second model is the 3-level HLM with students nested within schools nested within districts. This model allows us to calculate two ICCs, the school-level ICC and the district level ICC. The level-1, or student-level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad (3)$$

where y_{ijk} is the outcome for individual $i = \{1,..., n\}$ in school $j = \{1,..., J\}$ in district $k = \{1,..., K\}$; π_{0jk} is the average at school *j* in district *k*; and e_{ijk} is the residual error associated with each student.

The level-2 model, or school-level model is:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_{\pi}) \qquad (4)$$

where β_{00k} is the mean in district *k*, and τ_{π} is the variance between schools within districts.

The level-3 model, or district-level model is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad u_{00k} \sim N(0, \tau_\beta) \tag{5}$$

where γ_{000} is the grand mean and τ_β is the variance between districts. In the 3-level HLM, there are two ICCs. The between school within district ICC is $\rho_{L2} = \frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta}$ and the between district ICC is $\rho_{L3} = \frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta}$.

The third model is the within district 2-level HLM with students nested within schools. The unconditional models are the same as equation (1) and (2) except that they will be run within districts. A within district design is a common GRT which is why we are interested in examining within district ICCs.

For research question 1, we ran the three set of unconditional models for grades 5, 8, 10, and 11, the grades in which science is tested, for 5 years of data. We took the mean of the ICCs across the 5 years for each grade level.

For research question 2, we modified the unconditional models to include different covariates. We then calculated the R-square, or percent of variance explained at the relevant levels by the covariates. For example, the R-square at level 2 can be expressed as $R^2 = 1 - \frac{\tau_{\pi|X}}{\tau_\pi}$ where $\tau_{\pi|X}$ is the residual level-2 variance in a model with covariates and τ_π is the level-2 variance in the unconditional model. Due to space restrictions in this proposal, we do not provide the conditional models or R-square calculations. They are included in the full paper.

Table 1 depicts the grade level and years in which data is available. According to the unconditional row in Table 1, the models can be run at the 4 grade levels for all 5 years of data. The second row corresponds to the student demographics. The student demographics included dummy variables for female, black, Hispanic, and free/reduced lunch status. The models with student demographics are also available for all grade levels for all years as indicated in row 2. The next set of covariates included student-level pretest, which we defined as students' science score in a prior year. For grade 11, 1 year lags were available since students were also tested in science in grade 10. One year lags were not available for any other grade which is why grade 11 is the only grade listed in the row for 1-year lag. It is also important to note that the availability of the lag data depended on the year. For example, for grade 11, 1 year lag was available for the data from academic year 2, 3, 4, and 5 but not year 1. For grade 10, 2 year student lags were available, since students took the science test in grade 8. For grades 8 and 11, 3 year lags were also available since students took the test in grades 5 and 8, respectively. The next set of covariates included school means from the same grade lagged one, two, three, and four years. Note that these are available for all grades but not for all years since they require data from

earlier years. The last set of covariates included student and school lags. The available grades match the student lags since this data is more limited.

We also examined prior reading and math scores as covariates. Since math and reading tests are administered annually, all student and school level lags were available. The only limitation is in the year of the data. For example, 3 year school or student lags were only available for years 4 and 5 of the data.

Findings:

Table 2 presents the unconditional ICCs for the full data set for the 2-level HLM and 3-level HLM. Note that when district is included in the model, the between school variance is reduced by almost 70 percent from 0.174 to 0.054. The results from the within district 2-level HLM will be completed this fall.

The percent of variance explained by the covariate sets is documented in Table 3. Not all of the covariate sets have been run yet for all designs, so only a selection of models from the 2-Level design is presented. We find that in the 2-Level design, adding a one year lagged student pretest reduces the between schools variance by over 90 percent and the within schools variance by approximately 60 percent. As the duration of the lagged pretest is increased, both R-square values decline, as expected, but at a slow rate of less than 10 percent per year. In contrast, the addition of only demographic covariates explains only 50 percent of the between variance, and 10 percent of the within variance.

Conclusions:

The capacity of the field to conduct power analyses for GRTs of educational interventions has improved over the past decade (Authors, 2009). However, a power analysis depends on estimates of design parameters. Hence it is critical to build the empirical base of design parameters for GRTs across a variety of outcomes and contexts. This study provides a first step towards building this base of design parameters specifically for science outcomes. Unlike reading and math, science is not typically tested each year. Preliminary findings from this study suggest that although not direct comparisons, ICCs for science outcomes are smaller than grade 3 math and reading as reported by Bloom, Richburg-Hayes, and Black (2005; 2007) for five urban districts. Similarly, Hedges and Hedberg (2007) found larger ICCs for both reading and math using a nationally representative sample of students nested in schools. R-square values for school level covariates have not been computed yet, but a one year lagged student pretest appears to be highly effective in reducing variance between and within schools, more so than in reading (R-squares less than 0.86) and math (less than 0.63) as reported by Bloom et al. (2005; 2007). The empirical estimates from this study will help improve the accuracy of the power analyses for GRTs of science interventions.

Appendices

Appendix A. References

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). Using covariates to improve precision. MDRC working paper.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Donner, A. & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold Publishers.

Hedges, L., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal for Research on Educational Effectiveness, 3(2*), 157-198.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*, 265-288.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

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

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(*1*), 62-87.

Authors (2007).

Authors (2009).

Authors (2012).

Appendix B. Tables and Figures

Table 1. Grades in which data is available for models based on prior science scores.

	Year of Data				
	2006/07	2007/08	2008/09	2009/10	2010/11
Models	Grade level analysis is possible				
Unconditional	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
Student Demographics	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
1 year lag student scores science		11	11	11	11
2 years lag student scores science			10	10	10
3 years lag student scores science				8,11	8,11
1 year lag school level mean from science		5,8,10,11	5,8,10,11	5,8,10,11	5,8,10,11
2 years lag school level mean from science			5,8,10,11	5,8,10,11	5,8,10,11
3 years lag school level mean from science				5,8,10,11	5,8,10,11
4 years lag school level mean from science					5,8,10,11
1 year lag student science scores and school mean		11	11	11	11
2 years lag student science scores and school mean			10	10	10
3 years lag student science scores and school mean				8,11	8,11

Table 2. Average ICCs for unconditional model by grade, 2007-2011.

	2-Level (students in schools)	3-Level (students in schools in districts)		2-Level Within District (students in schools)
Grade	ICC	ICCL2	ICCL3	Average ICC
Grade 5	0.1906	0.0693	0.1186	Not Available Yet
Grade 8	0.1638	0.0547	0.1000	Not Available Yet
Grade 10	0.1762	0.0454	0.1256	Not Available Yet
Grade 11	0.1670	0.0486	0.1142	Not Available Yet
All Grades (Average)	0.1744	0.0545	0.1146	

Table 3. Average R-squared values for select conditional models by grade, 2007-2011

2-Level (students in schools)

Model	R^2	Grade 5	Grade 8	Grade 10	Grade 11	All Grades (Average)
Demographics	Between	0.5661	0.5962	0.5519	0.5470	0.5653
	Within	0.0688	0.1427	0.0864	0.0819	0.0949
Student Pretest (Lag 1)	Between	---	---	---	0.9303	0.9303
	Within	---	---	---	0.5983	0.5983
Student Pretest (Lag 2)	Between	---	---	0.8252	---	0.8252
	Within	---	---	0.5659	---	0.5659
Student Pretest (Lag 3)	Between	---	0.6663	---	0.8004	0.7334
	Within	---	0.4258	---	0.5238	0.4748

Note:
Individual grade averages for student pretest models are based on less than five years of data.