Testing Assumptions Behind the Analysis of Statistical Power: The Case of an RCT on Elementary Science Achievement

Jill Bowdon and Jeffrey Grigg
University of Wisconsin-Madison, Department of Sociology, Madison, WI

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant 144-NL14 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Introduction and Background

A power analysis determines the ability to discern a statistically significant difference in outcome-if one exists-between the treatment and control groups of an experiment. Conventionally, researchers assume that the power should be at least 0.80 and design their experiment to ensure that they will meet this threshold. This analysis examines how a randomized controlled trial (RCT) in the Los Angeles Unified School District made certain assumptions of power in its original grant application, and how these power assumptions needed to be modified based on the actual data.

The Data

Source: Data comes from a randomized controlled trial testing the effect of teacher development and student achievement in elementary science in the Los Angeles Unified School District. This NSF-funded study is called *"System-Wide Change: An Experimental Study of Teacher Development and Student Achievement in Elementary Science."* The primary investigators are Adam Gamoran and Geoffrey Borman.

The purpose of this randomized controlled trial is to test whether sustained, content-rich professional development in inquiry-based science instruction affects the science achievement of fourth and fifth graders.

Design: 10 schools from each of the 8 local districts in the Los Angeles Unified School District were randomly selected to participate. Of the 10, 5 were randomly assigned to treatment and 5 to control. Thus, randomization and analysis take place at the school level.

Treatment: Treatment schools (N=40) sent one or two teachers to participate in a summer training institute on science immersion. These teachers also received follow-up mentoring.

Control: Control school (N=40) received the immersion unit instructional guides but did not receive any professional development.

Outcome measures: Grade 5 standardized science achievement tests; Grade 4 and 5 periodic assessments

Research Problem and Methods

Power is Affected by:

Statistical power is affected by the magnitude of the treatment effect, the sample size, the intraclass correlation (ICC), alpha (the probability of a Type I error), and beta (the probability of a Type II error).

Initial Assumptions About Statistical Power for This RCT:

Assuming a minimum detectable effect size of 0.25; a sample size of 80 schools with 50 students in each school; an intraclass correlation (ICC) of 0.10, and an alpha of 0.05, *the power exceeds 0.80, even without the use of a covariate.*

An Unexpected Twist:

Using actual data from the 2005-06 school year and Hierarchical Linear Modeling (HLM) to partition the variance, we found that:

τ^2 = 563.98, σ^2 = 1816.95. Applying these values in the formula

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

gives us **an ICC of 0.24.**

This ICC is calculated using a sample size of 80 schools, and 108 students per cluster, as opposed to the originally anticipated 50 students per cluster.

Research Problem:

With this ICC, the power is at 0.611, far less than the 0.80 needed. The following analysis examines whether the use of a level two covariate can help increase the statistical power to an acceptable level.

Adding a Covariate to Help Estimate the Impact of the Treatment

For the covariate we chose to use a *school-level mean science test score for the year 2006*, which is a lagged outcome measure since the study's outcome measure is also a science achievement test.

Lagged outcome measures are the most powerful type of covariate because they capture all of the conditions that influenced the outcome before randomization and will likely influence the outcome in the future (Bloom 2005).

Methods:

We used Optimal Design Software to perform the power analysis (Liu, Spybrook, et al. 2006). We used HLM to calculate the covariate (Raudenbush, Bryk, et al. 2004).

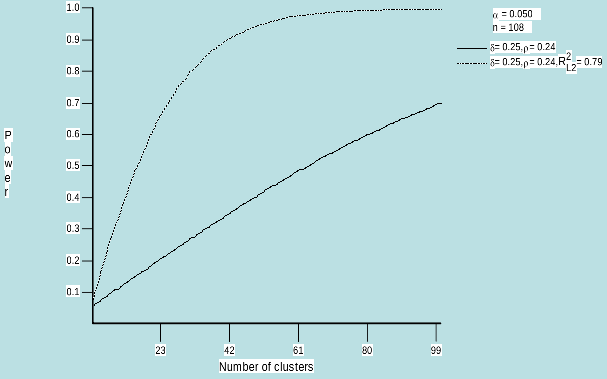
Results

The second level covariate, aggregate science achievement test scores, explained 79% of the variation between schools. This covariate raised the statistical power to detect an effect to 0.997. Now, researchers have a 99.7% chance of detecting a treatment effect if one exists. The jump in power is shown clearly in the table below.

TABLE: THE POWER INCREASES WITH THE ADDITION OF A COVARIATE

	Design 1 (without covariate)	Design 2 (With science covariate)
Minimum Detectable Effect Size (sigma)	.25	.25
Intra-Class Correlation (rho)	.24	.24
Probability of Type I Error (alpha)	.05	.05
Members per Cluster (n)	108	108
Number of Clusters (J)	80	80
Explained by Level-2 Covariate	0.00	0.79
Power	**0.611**	**0.997**

A GRAPHICAL REPRESENTATION OF POWER WITH AND WITHOUT A COVARIATE



Summary of Findings

Adding a level two measure as a covariate increased the statistical power, boosting it to an acceptable level. Now, instead of having only a 61.1% chance of detecting an impact of the treatment if one exists, researchers have a 99.7% chance.

This lagged outcome covariate soaked up variation between the schools because it controlled for the school and staffing effects that influenced science achievement before the randomization, and that continue to influence the outcome measure.

Conclusions and Future Directions

Adding a *cluster level covariate* helps control for some of the unexplained variation between schools. This reduces the cluster variance as a whole, improving the precision of the impact estimator.

We will continue to update the power analysis each year as we collect data. The intraclass correlation and the amount of variation explained by the level two covariate will fluctuate depending on the achievement tests from that year, therefore the power to detect an effect will also change.

References

Bloom, H. (2005). Randomizing Groups to Evaluate Place-Based Programs. Learning More from Social Experiments: Evolving Analytic Approaches. H. Bloom. New York, Russell Sage Foundation 115-172.

Liu, X., J. Spybrook, et al. (2006). Optimal Design For Multi-Level and Longitudinal Research

Raudenbush, Stephen, Anthony Bryk, Yuk Fai Cheong, Richard Congdon, and Mathilda du Toit. 2004. *HLM 6: Hierarchical Linear and Non Linear Modeling*. Lincolnwood: Social Scientific Software International, Inc.