Abstract Title Page

Title:
Strategies for Improving Power in Cluster Randomized Studies of Professional Development

Authors and Affiliations:
Ben Kelcey
University of Cincinnati

Jessaca Spybrook
Western Michigan University

Jiaqi Zhang
University of Cincinnati

Geoffrey Phelps
Education Testing Services

Nathan Jones
Boston University

Background / Context:
With research indicating substantial differences among teachers in terms of their effectiveness (Nye, Konstantopoulous, & Hedges, 2004), a major focus of recent research in education has been on improving teacher quality through professional development (Desimone, 2009; Institute of Educations Sciences [IES], 2012; Measures of Effective Teaching project [MET], 2012; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Notwithstanding widespread support for the development of teachers, there is a growing recognition of the lack of reliable empirical evidence concerning which features and programs of professional development are effective (Wayne et al., 2008). Consequently, there has been strong interest in supporting research that can inform the design of effective professional development programs (Desimone, 2009; IES, 2012; Wayne et al., 2008; Garet et al., 2011). For instance, through many different programs and topics, the Institute of Education Sciences (IES) has funded dozens of projects that targeted the professional development of teachers and has recently established an entire program devoted to research on effective strategies for improving teacher quality through professional development (IES, 2012).

Despite the national emphasis on improving teacher effectiveness and development, there has been little research discussing how to effectively design and implement teacher professional development studies (Wayne et al., 2008). Perhaps because of this lack of research, examples of professional development studies with high quality designs have been rare. A recent review of professional development studies found that less than one percent of studies sampled offered designs that would permit rigorous causal inference (Yoon et al., 2007). For these reasons, the field has called for more studies that evaluate the effectiveness of professional development programs on valued outcomes using rigorous designs (Barrett et al., 2012).

Purpose / Objective / Research Question / Focus of Study:
In this study, we empirically examined the comparative power and practical viability of several different types of cluster randomized trials in professional development studies. We outline why such designs are well suited for studies of many professional development programs. We then report estimates for parameters needed to plan such studies and use the estimates to explore the comparative efficiency of several designs. We examined three primary questions:
1) What is the variance decomposition of teacher knowledge outcomes across teachers, schools, and districts? The precision of treatment effect estimates and the statistical power of group randomized designs fundamentally depend on the variance decomposition across levels. Despite recent shifts in research and funding priorities emphasizing the value of carefully designed studies, research on study designs for the evaluation of professional development programs has lagged well behind its student outcome counterparts (e.g., Borko, 2004; Wayne et al., 2008; Yoon et al., 2007; Hedges & Hedberg, 2013). Our work aims to fill this gap by providing empirical estimates of the variance decomposition across levels for multiple teacher outcomes.
2) To what extent is there evidence that covariance adjustment on pretreatment covariates such as a pretest, teacher certification, or demographic covariates can reduce the sample size necessary to achieve a desired power level? An important conclusion from previous statistical and empirical studies of group randomized designs is that adjusting for differences on key covariates can substantially improve the power to detect treatment effects if they exist (Raudenbush, 1997). In many instances, the explanatory power of a pretest can be used to dramatically reduce the sample size necessary to adequately power a study and substantially lower the cost of the study (e.g., Bloom, 2005; Hedges & Hedberg, 2007). We examined the value of adjusting for a pretest as well as several school and teacher variables.

3) To what extent is there evidence that blocking on districts can reduce the sample size necessary to achieve a desired power level? Literature has demonstrated that multisite group randomized designs which assign treatments within blocks defined by hierarchical units can often be an effective strategy to reduce the sample sizes necessary to achieve a desired power level. We examined the value of designs which randomly assign intact schools within each district to treatment conditions and compared them to designs which randomly assign schools to treatments and ignore districts and designs which randomly assign intact districts.

Population / Participants / Subjects:
The data we report on in this proposal comes from over 10,000 teachers and 3000 schools, 200 districts and span five different teacher knowledge outcomes: (1) Elementary School Number Concepts and Operations, (2) Elementary School Patterns, Functions and Algebra, (3) Grade 4-8 Geometry, (4) Middle School Number Concepts and Operations and (5) Middle School Patterns, Functions and Algebra (e.g., Hill, Rowan & Ball, 2005; Hill et al., 2008).

Research Design:
A key feature of most professional development programs is that they are designed for and implemented by intact schools/districts because they are intended to promote and leverage social processes and learning (Borko, 2004). The active collaboration of teachers as they integrate professional development into their daily practice is often seen as critical features of effective professional development. Literature has consistently emphasized the importance of establishing study designs that align with the theories underlying programs being studied (e.g., 'theory-driven'; Chen & Rossi, 1983).

Single level designs which assign teachers within schools to different conditions have the potential to undercut the validity of study results because they must either eliminate collaboration altogether or allow collaboration across treatment conditions. Eliminating collaboration from the study may misrepresent or distort the theory of action underlying the effectiveness of the given professional development program because it suppresses the specific collaboration efforts among teachers that are thought to cultivate change. Similarly, allowing collaboration across treatment and control may introduce treatment diffusion because control teachers may receive some unknown portion of the treatment (Bloom, 2005). In turn, this diffusion obscures treatment-control contrasts and potentially violates the stable unit treatment value assumption. In this way, group-randomized designs are well suited for studies of professional development because they can accommodate programs that are delivered to intact groups (e.g., schools/districts), the collaborative nature of many professional development programs, and extant teacher/school assignments.

Significance / Novelty of study:
A principal consideration in professional development studies is the power with which a design can detect effects if they exist (e.g., Raudenbush, 1997). Though group designs may be theoretically favorable, prior research has suggested that they may be challenging to conduct in professional development studies because well-powered designs will typically require large sample sizes or expect large effect sizes. To assess and address these challenges, we report empirical evidence of the magnitude of clustering of teachers within schools and districts and the extent to which the efficiency of group randomized designs can be improved upon through covariance adjustment and/or blocking on districts.

Statistical, Measurement, or Econometric Model:

We focused on three particular designs which randomly assign interventions to (a) intact schools (disregarding districts), (b) intact districts, or (c) intact schools within blocks defined by districts. We examined the comparative performance of these designs under unconditional specifications and specifications that adjusted for the covariance of the outcome with pertinent covariates such as the pretest. The work we report in this proposal focuses on estimates of two parameters that are central to the planning and design of group randomized studies: intraclass correlation coefficients (ICCs) and variance explained at each level (R^2) (Raudenbush & Bryk, 2002).

Models. We estimated values of the design parameters using two and three level hierarchical linear models. For brevity we only outline the two level models which consider teachers nested within schools and ignore districts. We modeled the outcome, *Y*, for teacher *i* in school *j* as

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2) \\ \beta_{0j} &= \gamma_{00} + u_{0j} \qquad u_{0j} \sim N(0,\sigma_\beta^2) \end{aligned} \tag{1}$$

Here, β_{0j} is the school-specific intercept, ε_{ij} is the teacher-specific residual, γ_{00} is the grand mean and u_j is the random effect for school *j*. The unconditional ICC associated with this model was estimated as

$$\rho = \sigma_\beta^2 / (\sigma_\beta^2 + \sigma_\varepsilon^2) \tag{1}$$

To estimate the variance explained by the different covariate sets, we expand equation (1) to include covariates

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \sum_{p=1}^{P} \beta_{pj}(X_{pij} - \bar{X}_{pj}) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,\sigma_{A\varepsilon}^2) \\ \beta_{0j} &= \gamma_{00} + \sum_{p=1}^{P} \gamma_{0p}\bar{X}_{pj} + u_{0j} \qquad u_{0j} \sim N(0,\sigma_{A\beta}^2) \end{aligned} \tag{2}$$

with X_{pij} as covariate *p* for teacher *i* in school *j* (group-mean centered) with associated regression coefficient β_{pj} and $\bar{X}_{pj}$ to indicate the school level average with coefficient γ_{0p}. Residual variance, $\sigma_{A\varepsilon}^2$, and school variance, $\sigma_{A\beta}^2$, represent the adjusted variation in the outcome conditional upon the respective covariate set. Accordingly, we estimated the proportion of variance explained by a covariate set as

$$R_W^2 = (\sigma_\varepsilon^2 - \sigma_{A\varepsilon}^2)/\sigma_\varepsilon^2 \quad \text{and} \quad R_B^2 = (\sigma_\beta^2 - \sigma_{A\beta}^2)/\sigma_\beta^2 \tag{4}$$

at the teacher and school levels. Here $\sigma_\varepsilon^2, \sigma_{A\varepsilon}^2, \sigma_\beta^2$, and $\sigma_{A\beta}^2$ represent the unconditional teacher level variance, the conditional teacher level variance, the unconditional school-level variance, and the conditional school level variance, respectively. We examined multiple sets of covariates but for brevity only highlight models conditional on the pretest.

An important implication of the magnitude of the ICC and explanatory power of covariates is their impact on the sample size needed to adequately power a study to detect an intervention effect. A common way to summarize this implication is to estimate the minimum detectable effect size (MDES) for designs with varying sample sizes at each level (Bloom, 2005). Our work considered balanced unconditional designs as well as designs that conditioned on the covariates. For brevity, we outline MDESs for only two level school randomized designs which ignore district membership (but also consider 3 level designs in the full paper).We estimated the two level MDESs using

$$MDES = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\sigma_\beta^2(1-R_B^2)}{J} + \frac{\sigma_\varepsilon^2(1-R_W^2)}{I*J}} * \frac{1}{\sqrt{\sigma_\beta^2 + \sigma_\varepsilon^2}} \tag{5}$$

where *J* was the number of schools, *C* was the number of covariates, and *M* was a design multiplier such that, $M \approx t_{\alpha/2} + t_{1-\beta}$ for a *t* distribution with *J-2-C* degrees of freedom. Further, *P*

was the proportion of schools assigned to treatment, σ_β^2 and σ_ε^2 were the school and teacher level variances, R_B^2 and R_W^2 were the variance explained at the school and teacher levels.

Findings / Results:

Empirical Estimates. The variance decomposition results suggested substantial clustering among teachers within schools and within districts (Tables 1 and 2). Two level variance decompositions using hierarchical linear models with teacher nested within schools indicated that the variance attributable to differences among schools ranged from a low of 0.17 in the middle school number concepts and operations outcome to a high of 0.33 in the elementary geometry outcome (Table 1). Three level variance decompositions with teachers nested within schools nested within districts indicated that the variance attributable to (a) differences among schools ranged from a low of 0.03 in the elementary school number concepts and operations outcome to a high of 0.21 in the elementary geometry outcome and (b) differences among districts ranged from a low of 0.11 in the elementary school geometry outcome to a high of 0.18 in the elementary school patterns, functions, and algebra outcome (Table 2). Subsequent analyses suggested that a substantial portion of the variance attributable to each level was accounted for by teachers' prior abilities (Tables 3 and 4). Our results suggested that the precision of estimates and the statistical power of designs for studies of professional development with teacher knowledge outcomes will tend to be influenced by the clustering, covariance adjustment and blocking on districts.

Implications for Design. Using the estimates of the design parameters, we calculated MDES for different designs. For brevity we only present the MDES results for the unconditional designs for a few sample sizes but note that there are important differences in power under several conditional (e.g., on pretest) designs. A useful example is to choose a typical number of schools for a design and examine how MDESs vary. A typical sample in many studies using student outcomes is somewhere between 30 and 50 schools. To make our example concrete we estimated MDES for designs that use 10 districts with 3 or 5 schools per district and 4 or 10 teachers per school. Under the assumption that there are no district-by-treatment interactions and fixed effects are used to account for the district blocking, Table 5 displays the MDESs under several designs and sample sizes. Our results suggested that multisite school randomized designs tended to be the most efficient. On average, the multisite school randomized design tended to be able to detect effect sizes 20% smaller the school randomized designs (which ignore districts) and effect sizes 50% smaller than district randomized designs.

Conclusions:

A significant implication of the magnitude of the intraclass correlations for teachers in schools and districts is that group randomized studies of professional development interventions will tend to need large sample sizes or expect large effect sizes. However, the current work also highlights the efficacy of two established paths to improving the efficiency of study design and thereby reducing the sample sizes necessary to achieve a predetermined power level. Our results first suggested that covariance adjustment for prognostic covariates often provides substantial gains in efficiency. Similarly, our results suggested that blocking on districts improves efficiency. Collectively, the analyses suggested that covariance adjustment and/or blocking on districts has the potential to transform designs that are unreasonably large for professional development studies into viable studies. Despite the promising results, we are cautious to note that there are several limitations of our study including our consideration of only one type of professional development outcome and the nature of our sample. As a result, the extent to which the results might generalize to other samples or apply to other outcomes is unknown.

Appendices

Appendix A. References

Barrett, N., Butler, J. S., & Toma, E. F. (2012). Do less effective teachers choose professional development does it matter? *Evaluation review, 36*(5), 346-374.

Bloom, H. (2005). Randomizing groups to evaluate place-based programs. In H.S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York, NY: Russell Sage Foundation

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3–15.

Chen, H., & Rossi, P. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review*, *7*, 283-302.

Desimone, L., (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38,* 181-199.

Garet, M.S., Wayne, A.J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M.,. Brown, S., Hurlburt, S., Zhu, P., Sepanik, S & Doolittle, F. (2011). Middle School Mathematics professional development Impact Study: Findings After the Second Year of Implementation. U.S. Department of Education Report NCEE 2011-4024.

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

Hill, H.C., Rowan, B., & Ball, D. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal, Vol. 42,* 371–406.

Institute of Education Sciences. (2012). Research Grants Request for Applications for awards beginning in Fiscal Year 2013: CFDA Number 84.305A. U.S. Department of Education.

Measures of Effective Teaching [MET] (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Retrieved from the Bill and Melinda Gates Foundation Measures of Effective Teaching website: metproject.org

Nye, B., Konstantopoulous, S., & Hedges, L.V. (2004). How large are teacher effects? Education Evaluation and Policy Analysis, 26, 327-257.

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

Raudenbush, S. W. & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469–479.

Yoon, K.S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Appendix B. Tables and Figures

Table 1

Two Level Unconditional Variance Components and Confidence Intervals by Outcome

Outcome	ICC	Low	High
Grade 4-8 Geometry	.33	.27	.37
Elementary School Number Concepts & Operations	.19	.16	.21
Elementary School Patterns, Functions & Algebra	.29	.25	.32
Middle School Number Concepts & Operations	.17	.11	.22
Middle School Patterns, Functions & Algebra	.27	.24	.30

Note: ICC is intraclass correlation coefficient of teachers nested within schools (ignoring districts)

Low refers to the lower bound of the 95% bootstrapped confidence interval

High refers to the upper bound of the 95% bootstrapped confidence interval

Table 2

Three Level Unconditional Variance Components and Their Confidence Interval by Outcome

	Schools			Districts		
Outcome	ICC	Low	High	ICC	Low	High
Grade 4-8 Geometry	.21	.15	.25	.11	.03	.17
Elementary School Number Concepts & Operations	.03	.01	.05	.13	.10	.16
Elementary School Patterns, Functions & Algebra	.05	.02	.07	.18	.13	.22
Middle School Number Concepts & Operations	.03	.00	.09	.13	.07	.17
Middle School Patterns, Functions & Algebra	.14	.10	.17	.13	.09	.16

Note: ICC is intraclass correlation coefficient of teachers nested within schools and nested within districts

Low refers to the lower bound of the 95% bootstrapped confidence interval

High refers to the upper bound of the 95% bootstrapped confidence interval

Table 3

Proportion of Variance Explained by Pretest for Each Outcome for Two Level Models

Outcome	Teachers	Schools
Grade 4-8 Geometry	.09	.25
Elementary School Number Concepts & Operations	.05	.21
Elementary School Patterns, Functions & Algebra	.03	.14
Middle School Number Concepts & Operations	.09	.24
Middle School Patterns, Functions & Algebra	.06	.36

Table 4

Proportion of Variance Explained by Pretest for Each Outcome for Three Level Models

Outcome	Teachers	Schools	Districts
Grade 4-8 Geometry	.09	.32	.08
Elementary School Number Concepts & Operations	.05	.21	.16
Elementary School Patterns, Functions & Algebra	.03	.11	.10
Middle School Number Concepts & Operations	.09	.60	.22
Middle School Patterns, Functions & Algebra	.06	.46	.23

Table 5

Minimum Detectable Effect Sizes for Two Level School Randomized, Three Level District Randomized, and Multisite School Randomized Designs with Selected Sample Sizes

	Two Level School Randomized				Three Level District Randomized				Three Level Multisite (Blocked on Districts) School Randomized			
Districts	NA				10				10			
Schools	30		50		3/district		5/district		3/district		5/district	
Teachers/school	4	10	4	10	4	10	4	10	4	10	4	10
Grade 4-8 Geometry	0.75	0.67	0.57	0.51	0.98	0.91	0.87	0.82	0.67	0.57	0.50	0.43
Elementary School Number Concepts & Operations	0.66	0.55	0.51	0.42	0.93	0.83	0.85	0.79	0.53	0.37	0.40	0.28
Elementary School Patterns, Functions & Algebra	0.72	0.64	0.55	0.49	1.03	0.95	0.97	0.92	0.53	0.39	0.40	0.30
Middle School Number Concepts & Operations	0.65	0.53	0.5	0.41	0.93	0.83	0.85	0.79	0.53	0.37	0.40	0.28
Middle School Patterns, Functions & Algebra	0.71	0.62	0.54	0.47	0.84	0.73	0.76	0.68	0.61	0.50	0.46	0.38