Abstract Title Page

Title:
Parameters for the Design of Group Randomized Studies for Teacher Professional Development

Authors and Affiliations:

Ben Kelcey
Wayne State University
ben.kelcey@gmail.com

Geoffrey Phelps
Education Testing Services
gphelps@ets.org

Nathan Jones
Education Testing Services
ndjones@ets.org

Background / Context:

Teacher professional development (PD) is seen as critical to improving the quality of US schools (National Commission on Teaching and America's Future, 1997). PD is increasingly viewed as one of the primary levers for improving teaching quality and ultimately student achievement (Correnti, 2007). One factor that is driving interest in PD is rapidly mounting evidence that teachers vary greatly in their effectiveness (Nye et al., 2004). Another is a growing recognition that there is much for teachers to learn if they are to realize ambitious new content standards for students (Ball & Cohen, 1999; Borko, 2004). To this end, policymakers and funding agencies have directed considerable resources toward PD studies focused on improving teacher effectiveness (Birman et al. 2007).

However, many argue that the current system of PD is not up to the task. By most accounts, PD is fragmented and dominated by district-led workshops that are typically short in duration, disconnected from teachers' work experiences, and seldom focused on the daily challenges of teaching core subjects such as reading and mathematics (Ball & Cohen, 1999; Wilson & Berne, 1999). While there are many examples of more substantive PD, these are typically offered through universities, taught by the developers of the programs, and not available at sufficient scale (Hill, 2009).

A growing recognition of both the challenges and importance of improving PD has led to widespread interest in supporting research that can inform the design of effective programs. However, the existing empirical research base has been limited to mostly small-scale evaluations of single-site programs with questionable outcome measures and designs that are poorly suited to drawing causal inferences (Borko, 2004; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). For instance, studies often simply ask teachers participating in a program how much they think they have learned or the extent to which they found the program useful for improving their teaching (Wilson & Berne, 1999). Similarly, recent reviews of PD studies indicated that only 9 of about 1,300 studies offered research designs that would permit causal inference (Yoon et al. 2007). Only a handful of studies were designed to investigate the impact of PD programs with rigorous designs and on valued and appropriate outcomes (Desimone, et al., 2002).

What components of a study design would allow for more rigorously examining the effectiveness of PD? Desimone (2009) suggested a basic model for studying PD that includes systematic attention to three critical outcomes of PD: teacher knowledge, teaching quality, and student learning. Attending to each of these features in more systematic studies would allow for testing a "theory of teacher change (e.g., that PD alters teacher knowledge, beliefs, or practice) and a theory of instruction (e.g., that changed practice influences student achievement), both of which are necessary to complete our understanding of how PD works" (Desimone, 2009, p. 185). This basic model highlights the central importance of teacher knowledge. Change in teacher knowledge is the most direct and proximal outcome of PD and it is also an important mediator that could explain effects of PD on student outcomes.

There is empirical evidence supporting teachers' knowledge as a critical ingredient and central outcome of effective PD. Recent research has linked PD with changes in teachers' knowledge and quality and student achievement (Correnti, 2007). Further, Kennedy (1999) found that "programs whose content focused mainly on teachers' behaviors demonstrated smaller influences on student learning than did programs whose content focused on teachers' knowledge..." (p. 17). More recent literature has also established links among teacher knowledge and student learning in multiple subjects (Carlisle, Kelcey, Phelps & Rowan, 2011).

Purpose / Objective / Research Question / Focus of Study:

Despite recent shifts in research emphasizing the value of carefully designed experiments, the number of studies of teacher PD with rigorous designs and outcomes has lagged behind its student outcome counterparts. The purpose of the work reported in this proposal is to provide guidance to the research and policy community on the design of rigorous studies that have the power to draw strong causal inferences to establish the effectiveness of PD programs on valued outcomes. Our work considered a new generation of mathematics and reading knowledge outcomes designed to assess the types of content problems that teachers encounter in practice. Using these outcomes, we developed empirical estimates of design parameters such as the intraclass correlation coefficient (ICC) for teacher PD studies. Within this framework we first investigated the extent to which teachers' knowledge levels were clustered in schools and the extent to which this clustering varies by outcome domain and school context. Second, and aligned with the conference theme, we investigated the extent to which uses of general design parameter estimates (e.g., ICCs irrespective of outcome) leads to indeterminacies in design. Specifically, a key finding that has emerged from research on educational outcomes is that design parameters vary considerably across outcomes and that a study's power is sensitive to these differences (Jacob et al., 2010). To investigate this sensitivity, we examined the extent to which the appropriateness of design parameters in mounting a well powered study were sensitive to the knowledge domain and the contextual features of studies. We then provide empirical estimates of domain- and context-specific design parameters to help researchers navigate methodological choices that lead to more effective PD study design.

Setting and Population / Participants / Subjects:

The current study comes out of two studies designed to develop new measures of teachers' knowledge in math and reading--the Learning Mathematics for Teaching and the Assessment of Reading Knowledge projects. Measured knowledge items have been constructed from large pools of questions developed over the course of the projects and are intended to provide measures useful across a wide range of math and reading education research projects. In this proposal we highlight 6 different elementary and middle school outcomes: (1) Elementary number and operations; (2) Elementary patterns, functions and algebra; (3) grade 4-8 geometry; (4) Middle school number and operations; (5) Middle school algebra; and (6) early elementary reading. Data for the mathematics outcomes come from 41 states and the District of Columbia whereas reading data are drawn only from a single state. Our sample represents teachers who participated in the Reading First program or PD programs, and programs varied with regard to whether teachers were required (or volunteered) to participate. The sample includes over 4,000 teachers representing over 500 schools. While the data are not necessarily representative of teachers from these states, the Reading First program or the larger U.S., they represent one of the largest samples of teacher PD programs to date. We present brief descriptives in Tables 1 and 2.

Research Design:

A key feature of most PD programs is that they are designed for and implemented by intact schools/districts. The active involvement of collaborative groups of teachers as they integrate PD into their daily practice is intended to promote and leverage social processes and learning (Borko, 2004). While statistically it is convenient favorable to assign individual teachers within the same school to different treatments, doing so removes a critical feature of contemporary PD. To this end, the work we present considers designs which randomly assign interventions to intact groups (e.g., teachers nested within schools). For each of the aforementioned outcomes we examined unconditional variance decompositions, conditional

variance decompositions, variance explained (pseudo-R^2) by multiple covariates and how each of these vary by outcome and several contextual features of a sample (e.g., urbanicity). Due to space constraints, however, our proposal only discusses the unconditional ICC results and only considers how these ICCs vary by outcome, grade and urbanicity.

Significance / Novelty of study:

To help improve the planning and design of GRTs, recent studies have compiled a wide range of empirical ICCs to describe the variance decomposition of many educational outcomes such as academic and behavioral indicators (Hedges & Hedberg, 2007; Jacobs, Zhu & Bloom, 2010). However, the scope of this literature has been limited to design parameters for student outcomes. In designing GRTs for PD, some researchers have either ignored clustering or used student parameters on the assumption that the nesting structure for students and teachers within schools are similar. Yet, a key finding that has emerged from research on educational outcomes is that design parameters vary considerably across outcomes and that a study's power is sensitive to these differences (Jacob et al., 2010).While the assumptions underlying the similarity of student and teacher design parameters are suspect, lacking suitable estimates of ICCs for teacher knowledge outcomes, researchers have few other options than to draw from student estimates. To this end, the work we present is designed to address these issues.

Statistical, Measurement, or Econometric Model:

To estimate ICCs and the variance explained by the above covariate blocks, we used a random intercept hierarchical linear model

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \sum_{p=1}^{P} \beta_{pj} X_{pij} + \varepsilon_{ij} \\ \beta_{0j} &= \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} \bar{X}_{qj} + u_j \end{aligned} \tag{0.1}$$

where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $u_j \sim N(0, \sigma_u^2)$. We estimated ICCs using

$$ICC = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2) \tag{0.2}$$

Findings / Results:

Overall, the average unconditional ICC across all outcomes was 0.18 suggesting that there was considerable clustering among teachers within schools (Tables 3-4). However, the magnitudes of the ICCs were highly sensitive to the outcome domain and sample context. Take for instance the differences among math outcomes. Although on average 21% of the variation in teachers' math knowledge was attributable to schools, only 13% was attributable to schools for the Elementary School Patterns, Functions & Algebra outcome while as much as 30% of this variation was attributable to schools for the grade 4-8 geometry outcome (Table 3). Similarly, while the overall ICC for the reading outcome across all three grades was 0.13, ICCs for individual grades ranges from a high of 0.24 in grade one to a low of 0.08 in grade three. The sensitivity of design parameters also extended to the contextual differences among schools. We present one simple example of context by describing how ICCs vary between urban and rural schools (Table 4). Evident from this table, ICCs differ considerably between urban and rural schools but even these differences are further dependent on the outcome domain.

Usefulness / Applicability of Method:

To illustrate the importance of the appropriate values for design parameters in designing group randomized PD studies and to put our results into context, consider several recently funded studies. First, consider a study which focused on the impact of early reading PD

interventions on the knowledge and practice of teachers in second grade (Garet et al., 2008). The study randomly assigned 90 intact schools (with approximately 3 teachers per school) to treatment conditions. With a 0.80 power level and a type one error rate of 0.05, the minimum detectable effect size (MDES) (Bloom, 2005) for an ICC 0.08 (i.e., unconditional ICC for reading knowledge in grade 2) is approximately 0.37. Results of the PD study found a statistically significant difference between treatment groups with an effect size of about 0.37 (Garet et al., 2008). The results of this study in conjunction with our analyses would suggest that the study design was well powered for the magnitude of the effects.

Now consider the Middle School Mathematics PD Impact Study (Garet et al., 2011). The study used a school randomized design with 39 schools and about 3 teachers per school to examine the extent to which specialized PD on rational numbers changed teachers' knowledge. Researchers reported that the specialized PD program did not have a statistically significant impact on teachers' knowledge of rational numbers and that the observed effect size was quite small (Garet et al., 2011). Our results would suggest that the MDES for this outcome and design would be on the order of 0.70 (0.60 with useful covariates). In other words, this study design was likely only sufficient for detecting large program effects.

To further illustrate the importance of the differences among ICCs in designing studies of PD, consider the differences in power for researchers trying to design a study to detect a treatment effect of size 0.50 with a reading knowledge outcome for grade one.
With four teachers per school, adequate power (0.80) for an ICC of 0 would only require 34 groups. In contrast, using the general ICC estimate for this outcome (irrespective of grade, 0.13) would indicate that adequate power would require a 30% large sample size--45 schools. However, our results question whether even this adjustment is sufficient because the grade one-specific ICC estimate (0.24) suggests that adequate power can only be achieved with 57 schools. Put differently, if the true ICC was 0.24, studies designed with adequate power for an ICC of 0 or 0.18 would retain only a 57% or 69% chance of detecting the treatment effect respectively.

Conclusions:

Collectively, the results underscored the importance of appropriate values for design parameters in planning group randomized studies in teacher PD. Our results also suggested that there was substantial variation in the observed ICCs across subjects, domains, grades and urbanicity. These findings suggest the need to establish outcome- and context- specific empirical benchmarks. However, we are cautious to not that although there was substantial variability among the estimated ICCs, these observed differences could be due to sampling error and/or generalizability error. Sampling error stems from the chance differences between the sample and population of interest. Our estimates yielded interval widths of about 0.20 suggesting significant overlap in outcome- and context-specific ICC confidence intervals and suggest that outcomes could have shared similar ICCs. Studies repeating our work would help to explicate the extent to which ICCs differ by outcome and context. In contrast to sampling error, generalizability error manifests when there is a true underlying difference between a target population and reference sample (Jacob, Zhu & Bloom, 2010). This issue may be relevant for the values we report for two reasons. First, our data were not representative samples of any specific populations and, as a result, it is unclear how the characteristics of this sample might relate to schools nationally, regionally or within other types of stratification. Second, while our overall sample size is likely the largest to date of such measures, it is still relatively small when compared to similar student level databases (e.g., Hedges & Hedberg, 2007). As a result, it is unclear how the characteristics of this sample might relate to schools nationally, regionally or within other types of stratification.

Appendices

Appendix A. References

Borko, H. (2006). American Educational Research Association annual presidential address.

Carlisle, J., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' Knowledge About Early Reading: Effects on Students' Gains in Reading Achievement. *Journal for Research on Educational Effectiveness,* 4, pp. 289-321

Desimone, L. (2009). Improving impact studies of teachers' PD: toward better conceptualizations and measures. *Educational Researcher,* v. 38, 3, pp.181-199.

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.

Hill, H. C. (2001). Policy Is Not Enough: Language and the Interpretation of State Standards. *American Educational Research Journal, 38*(2), 289-318.

National Mathematics Panel (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: U.S. Department of Education.

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

Spybrook, J. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis,* 31, 3, pp. 298-318.

Appendix B. Tables and Figures

Table 1: Selected descriptive statistics for the Investigating the Development and Measurement of Mathematical Knowledge for Teaching study

Variable	Mean	SD
Teacher characteristics		
Number of mathematics courses	4.17	2.95
Number of mathematics teaching method courses	1.98	2.00
Majored/minored in mathematics	0.19	0.39
Male	0.13	0.34
White	0.80	0.40
School characteristics		
Proportion of students who are White	0.51	0.34
Title 1 school	0.80	0.40
Proportion of students who are eligible for free/reduced lunch	0.57	0.24

Table 2: Selected descriptive statistics for the Assessment of Reading Knowledge study

Variable	Mean	SD
Teacher characteristics		
Number of reading courses	3.49	1.58
Majored/minored in reading	0.12	0.32
Certified in reading	0.64	0.48
Male	0.07	0.26
White	0.81	0.39
School characteristics		
Proportion of students who are White	0.35	0.35
Title 1 school	1.00	--
Proportion of students who are eligible for free/reduced lunch	0.77	0.19

Table 3: Unconditional ICCs and their confidence intervals

Model	*N*	*J*	ICC	Low[a]	High[b]
Mathematics					
Unconditional, Elementary School Number Concepts & Operations	795	385	0.15	0.07	0.24
Unconditional, Elementary School Patterns, Functions & Algebra	562	250	0.13	0.03	0.24
Unconditional, Grade 4-8 Geometry	492	270	0.30	0.17	0.41
Unconditional, Middle School Number Concepts & Operations	243	171	0.18	0.00	0.36
Unconditional, Middle School Patterns, Functions & Algebra	708	446	0.27	0.14	0.38
Average ICC of five math outcomes	--	--	0.21	--	--
Reading					
Unconditional, Grade 1	318	125	0.24	0.11	0.36
Unconditional, Grade 2	301	123	0.08	0.01	0.21
Unconditional, Grade 3	326	129	0.12	0.01	0.25
Unconditional, Grades one to three	947	129	0.13	0.06	0.21
Overall Average ICC of all outcomes	--	--	0.18	--	--

[a]Low refers to the lower bound of a 95% confidence interval based on multilevel bootstrap
[b]High refers to the upper bound of a 95% confidence interval based on multilevel bootstrap
Note: N refers to the teacher sample size and *J* refers to the school sample size. Columns describe the teacher sample size, the school sample size, the estimated ICC, and the lower and upper bound of the 95% confidence interval for the ICC. Rows pertaining to mathematics outcomes specify the covariate set and the outcome. Rows pertaining to reading outcomes specify the covariate set and the grade level.

Table 4: Comparison of unconditional ICCs among urban, rural and all[a] schools

	Urban	Rural	All
Elementary School Number Concepts & Operations	0.07 (*N*=407, *J*=170)	0.24 (*N*=359, *J*=186)	0.15 (*N*=795, *J*=385)
Elementary School Patterns, Functions & Algebra	0.03 (*N*=205, *J*=84)	0.24 (*N*=338, *J*=149)	0.13 (*N*=562, *J*=250)
Grade 4-8 Geometry	0.25 (*N*=265, *J*=118)	0.31 (*N*=212, *J*=137)	0.30 (*N*=492, *J*=270)
Middle School Patterns, Functions & Algebra	0.25 (*N*=314, *J*=178)	0.28 (*N*=362, *J*=238)	0.27 (*N*=708, *J*=446)
Grade 1-3 Reading and Reading Practice	0.12 (*N*=468, *J*=76)	-	0.13 (*N*=936, *J*=151)

[a]All schools refers to urban, rural and suburban schools
Note: ICCs were not estimated for outcome-urbanicity combinations that had less than 50 teachers
Note: *N* refers to teacher sample size and *J* refers to the school sample size

Figure 1: Power as a function of the number of clusters for a cluster randomized trial with 4 units per cluster, type 1 error rate of 0.05, effect size of 0.5 and intraclass correlations of 0, 0.13 and 0.24