THE JOURNAL OF EXPERIMENTAL EDUCATION
http://dx.doi.org/10.1080/00220973.2016.1220911

Routledge
Taylor & Francis Group

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN

Designing Large-Scale Multisite and Cluster-Randomized Studies of Professional Development

Ben Kelcey[a], Jessaca Spybrook[b], Geoffrey Phelps[c], Nathan Jones[d], and Jiaqi Zhang[a]

[a]University of Cincinnati, Cincinnati, Ohio; [b]Western Michigan University, Kalamazoo, Michigan; [c]Educational Testing Service, Princeton, New Jersey; [d]Boston University, Boston, Massachusetts

ABSTRACT

We develop a theoretical and empirical basis for the design of teacher professional development studies. We build on previous work by (a) developing estimates of intraclass correlation coefficients for teacher outcomes using two- and three-level data structures, (b) developing estimates of the variance explained by covariates, and (c) modifying the conventional optimal design framework to include differential covariate costs so as to capture the point at which the cost of collecting a covariate overtakes the reduction in variance it supplies. We illustrate the use of these estimates to explore the absolute and relative sensitivity of multilevel designs in teacher professional development studies. The results from these analyses are intended to guide researchers in making more-informed decisions about the tradeoffs and considerations involved in selecting study designs for assessing the impacts of professional development programs.

KEYWORDS

Cluster-randomization trials; experimental design; multisite cluster randomized trials; professional development

RECENT EFFORTS TO improve research in education have emphasized the value of well-designed studies and documented specific design strategies to improve the quality and scope of inferences resulting from studies in education (e.g., Spybrook & Raudenbush, 2009). Although these strategies have been widely developed and implemented in education studies for student outcomes, their development and use in professional development studies with critical intermediate outcomes has trailed behind significantly. Yet teacher professional development is seen as critical to improving the quality of schooling and student growth because there is mounting evidence that teachers vary greatly in their effectiveness (e.g., Nye et al., 2004; Darling-Hammond, 1997). Recognition of the challenges and importance of improving teaching has led to widespread interest in research that can inform the rigorous design of professional development programs (e.g., Wayne et al., 2008; Kennedy, 2016).

Policymakers and funding agencies have directed considerable resources toward professional development studies focused on improving teacher effectiveness through teacher development (Birman et al., 2007). For instance, through many different programs and topics, the Institute of Education Sciences (IES) has funded scores of professional development and teacher effectiveness projects. With the goal of developing an empirical basis for promoting and developing effective teaching, IES has recently established an entire program devoted to research on effective teaching and teacher development (IES, 2012).

Despite the national emphasis on improving teaching effectiveness, there have been several significant issues hampering rigorous research on teacher professional development. A critical issue is that although there is an empirical basis concerning the types of knowledge teachers need (e.g., Hill, Rowan, & Ball, 2005), literature delineating theories of how teachers learn and update their practice are

CONTACT Ben Kelcey ben.kelcey@gmail.com
© 2016 Taylor & Francis

underdeveloped (Kennedy, 2016). An important deficit in this area is the lack of clarity about the mechanisms through which professional development operates.

Although the theories of action underlying programs vary widely, research has typically outlined the general professional development process as it relates to student achievement by focusing on three sequential and increasingly distal outcomes of professional development: teacher knowledge, classroom instruction, and student learning (Desimone, 2009). Professional development programs are generally designed to build teacher knowledge and skills, which in turn are thought to improve classroom teaching and effect gains in student achievement (Yoon et al., 2007). The comprehensive study of each of these outcomes is critical to understanding how professional development programs come to be effective and mapping the teacher development process (Desimone, 2009).

Assessments of the mechanisms through which professional development operate are critical to assessing the validity of and advancing scientific theories underlying teacher development (Kennedy, 2016). Black box approaches that only estimate the impact of a program on, for example, student achievement are limited because the results can inform us of whether the program improved achievement but cannot explain why or provide valid assessments of the theory of action that underlies the program. For these reasons, proximal outcomes like teacher knowledge and instruction often represent critical outcomes of professional development and mediators through which we can test the theory of action.

A second complementary methodological issue is clarifying the types of designs that are amenable to the theory of contemporary professional development (Kennedy, 2016; Wayne et al., 2008; Yoon, Zhu, Cronen, & Garet, 2008). Many teacher development programs are structured to foster collaborative social processes, learning, and instruction and thus are designed to operate at the school- or district-level (Blank et al., 2008; Borko, 2004; Desimone, Garet, Birman, Porter, & Yoon, 2003; Kennedy, 2016). Literature has consistently emphasized the importance of aligning the design of a study with the theory underlying the intervention (e.g., Rossi, Lipsey, & Freeman, 2004).

Many professional development programs rely on collective participation and collaboration among teachers within schools or districts as a critical lever to promote teacher change (Kennedy, 2016). For these programs, professional development necessarily integrates learning opportunities across teachers in schools and/or a district. Study designs that directly incorporate consideration of the multilevel structure of professional development and schooling permit a more faithful implementation and representation of the collaboration that undergirds many theories of contemporary professional development.

Cluster-randomized and multisite (blocked) cluster-randomized designs may be well-positioned to improve the evidentiary base surrounding effective professional development. Cluster-randomized and multisite cluster-randomized designs have played a critical role in evaluating a broad array of educational initiatives (e.g., Spybrook & Raudenbush, 2009; Spybrook, Shi, & Kelcey, 2016). When implemented correctly such designs can accommodate the precise types of multilevel theories of action, structures, and treatment assignments that are common in many professional development programs.

At the same time, though potentially better aligned with many professional development programs in theory, cluster-randomized and multisite cluster-randomized designs may be impractical in some settings. For instance, such designs tend to be less efficient as compared to individually randomized designs, often requiring larger sample sizes to achieve comparable power. For this reason, there has been substantial research into accounting for the reduced efficiency through *a priori* estimates of the intraclass correlation coefficient for a wide variety of valued outcomes and populations (e.g., Spybrook, 2013).

For instance, empirical investigations of the intraclass correlation coefficient and their impacts on design have been documented for student outcomes such as mathematics (e.g., Bloom, Richburg-Hayes, & Black, 2005; Hedges & Hedberg, 2007a; Konstantopoulos, 2009), writing (Jaciw, Lin, & Ma, 2014), verbal fluency (Jacob, Zhu, & Bloom, 2010), reading comprehension (Brandon, Harrison, & Lawton, 2013; Hedberg & Hedges, 2014; Schochet, 2005), science (Spybrook, Westine, & Taylor, 2016; Westine, Spybrook, & Taylor, 2014), attendance (Jacob, Zhu, & Bloom, 2010), emotional and behavioral outcomes (Dong et al., in review; Jacob, Zhu, & Bloom, 2010), adult basic education outcomes

(Bloom, 1995), and health outcomes (Jacob, Zhu, & Bloom, 2010). Similarly, this research has been applied to a wide range of student populations including pre-kindergarten grades (e.g., Schochet, 2005), elementary grades (e.g., Schochet, 2008), and high school grades (e.g., Spybrook, Westine, & Taylor, 2016); rural and urban populations (e.g., Hedges & Hedberg, 2007b); U.S. populations (Hedberg & Hedges, 2014), and European and global populations (Zopluoglu, 2012).

The foundational value of this type of research has also been well established outside the boundaries of education and has included similar work on, for example, agricultural outcomes (e.g., Geyer, in review), epidemiological outcomes (Isaakidis & Ioannidis, 2003), HIV and pregnancy prevention outcomes (Glassman, Potter, Baumler, & Coyle, 2015), and household economic outcomes (Geyer, in review). Collectively, this research has led to the formation of empirically based guidelines that provide a careful mapping between study designs involving these outcomes, the organization of schooling or the multilevel structure, and the nature of the intervention under study. In turn, these guidelines have been correlated with dramatic improvements in study design and capacity over recent decades and have been identified as a critical step toward improving the field's capacity to conduct high-quality research studies that support causal inferences (Spybrook, Shi, & Kelcey, 2016; Spybrook & Raudenbush, 2009).

Within education, however, this line of inquiry has almost exclusively focused on study design using student outcomes. There has been very little research establishing an empirical basis and probing the feasibility of multilevel designs within the context of teacher professional development. Yet, critical assessments of the quality and rigor associated with teacher development studies have demonstrated that very few studies leverage designs that buttress causal inferences and provide adequate statistical power to detect non-negligible effects (Kelcey & Phelps, 2013; Kennedy, 2016; Yoon et al., 2007). Understanding the types of designs that are practically and theoretically amenable to studies of contemporary professional development programs is exceptionally important because it directly influences the scope of data samples and controls the types of research questions we can address with multilevel designs (Schochet, 2011).

Collectively, the results from previous studies outside of professional development point toward a pressing need for empirical estimates of design parameters specific to professional development studies (e.g., Jacob, Zhu, & Bloom, 2010). In particular, an important outcome of previous literature concerning the magnitudes of the intraclass correlation and variance explained parameters is that values vary widely both within a discipline across outcomes and across disciplines (e.g., Spybrook, 2013). For instance, within a single study, Jacob, Zhu, and Bloom (2010) reported intraclass correlation coefficients ranging from zero to over 0.30 across a dozen different student outcomes. These observed differences have important implications for study design because the sensitivity of multilevel designs to detect effects is inversely proportional to the intraclass correlation coefficient. More practically, power estimates tend to be susceptible to disparities between the assumed and actual values of the intraclass correlation coefficient (e.g., Hedges & Hedberg, 2007; Spybrook & Raudenbush, 2009). Given the national emphasis on teacher effectiveness and professional development and the increasing number of large-scale nationally funded professional development studies, it is critical that we develop the resources and capacity to adequately design professional development studies and understand their capacity to address causal questions of policy interest (IES, 2012; Spybrook, Shi, & Kelcey, 2016).

Purpose

In this study, we build on previous work on the design of professional development studies through three complementary aims. First, using a large national data set on teacher professional development, we develop estimates of intraclass correlation coefficients for teacher knowledge outcomes using two- and three-level data structures. The careful planning of cluster- and multisite cluster-randomized studies requires information concerning plausible values of the intraclass correlation coefficients specific to the outcome. We provide empirical estimates of the variance decomposition across teachers, schools, and districts for teacher knowledge outcomes.

Second, we estimate the variance in these outcomes explained by several covariates under the same two- and three-level data structures. An important conclusion from previous statistical and empirical studies of design efficiency is that adjusting for differences on key covariates can substantially improve the power and feasibility of designs. In many instances, the explanatory power of covariates, such as the pretest, can be used to reduce the sample size necessary to adequately power a study and thus substantially lower the cost of the study (e.g., Hedges & Hedberg, 2007). We study the value of adjusting for a pretest as well as several other variables across multiple outcomes and a wide range of professional development programs.

Finally, we use the aforementioned empirical estimates to explore the absolute and relative sensitivity of several types of cluster- and multisite cluster-randomized designs in terms of the minimum detectable effect sizes they are likely to yield in professional development studies. The analyses outline how one might use the plausible estimates to evaluate and improve designs through three connected strategies. First, we illustrate how to map out the relative sensitivity (as measured by the minimum detectable effect size) of several designs for an assumed cost structure. Second, we introduce a strategy to examine when and to what extent designs that use covariance adjustment outperform unconditional designs. An important practical limitation of extant literature concerning study design is that it generally does not consider the tradeoff between the additional costs associated with collecting covariate data and the resulting decrease in sample size under a fixed budget. For example, although covariance adjustment on prognostic covariates may produce more-efficient designs, collecting covariates is often paired with increased costs. Given a fixed budget, such increased costs may result in smaller sample sizes and ultimately a less sensitive design in terms of the minimum detectable effect size. We develop a simple strategy based on the conventional optimal design framework to examine tradeoffs associated with alternative (un)conditional design specifications with different costs.

Method

Data

To explore empirical estimates of design parameters as they apply to the design of multilevel professional development studies, we drew on data from the Investigating the Development and Measurement of Mathematical Knowledge for Teaching study. This study surveyed teachers across the United States as they participated in professional development programs. Teachers were measured in the specific knowledge area associated with the content of their professional development before and after they participated in a professional development program. Overall the sample contained 152 different professional development programs. These programs were diverse and included, for instance, trainings that varied in length, depth, and density (e.g., lasting from a few days to multiple years), varied in pedagogical focus (e.g., content knowledge, instructional practice), and varied in substantive content focus (e.g., geometry, algebra).

Despite its scale, our sample was a convenience sample of teachers, schools, districts, and professional development programs. Professional development providers self-selected the mathematical knowledge for teaching assessment and in many programs teachers may have self-selected to participate in the professional development program. It is also unclear how the characteristics of this sample might relate to teachers or schools nationally, regionally, or along other types of stratification. Further, we note that empirical estimates of variance and covariance can be susceptible to systematic differences among subgroups, features of professional development programs, and types and areas of outcomes. As a result, we are cautious to note that our subsequent analyses serve only as a preliminary investigation and illustration of design considerations for studies of professional development.

Across the outcomes, our samples represented 9,352 teachers, 3,540 schools, and 1,586 districts, with sample cluster sizes generally ranging from two to eight teachers per school and two to four schools per district. Overall, just over a fifth of the teachers maintained a certification for teaching mathematics and the majority had between 4 and 15 years of experience. Table 1 presents summaries of the teachers.

Table 1. Sample descriptive statistics.

	Sample Size						Years of Experience		
Outcome	Teachers	Schools	Districts	Pre-test	Post-test	Math Certified	0–3	4–15	>15
EG	727	348	147	.24 (.92)	.48 (.87)	.23	.16	.58	.26
ESNCO	3,337	992	371	.02 (.87)	.29 (.90)	.07	.20	.53	.27
ESPFA	2,071	610	237	−.18 (.95)	.02 (1.07)	.07	.19	.55	.26
MSNCO	1,019	502	266	.19 (.89)	.38 (.98)	.37	.25	.54	.21
MSPFA	2,198	1,088	565	.15 (.93)	.26 (.92)	.41	.23	.57	.20

Notes. ESNCO: elementary school number concepts and operations; ESPFA: elementary school patterns, functions, and algebra; EG: elementary geometry; MSNCO: middle school number concepts and operations; and MSPFA: middle school patterns, functions, and algebra. We present the mean for categorical variables and the mean and standard deviation for continuous variables.

Measures

Our analyses in this study focused on assessments drawn from the Mathematical Knowledge for Teaching (MKT) scales developed by the Learning Mathematics for Teaching project (http://www.umich.edu/~lmtweb/; Hill, Schilling, & Ball, 2004). MKT items represent one of the most fully developed examples of content knowledge for teaching assessments. The items were designed to focus directly on the particular types of mathematics knowledge teachers use in mathematics teaching. In contrast to conventional content questions that simply assess whether teachers themselves can do the mathematics that students are expected to learn, MKT measures focus on the specialized forms of mathematical knowledge only encountered in teaching mathematics. The assessment tasks focus on the mathematical problems that teachers encounter as they, for example, look for patterns in student errors, size up whether a nonstandard approach would work in general, or select a content representation that is well suited to supporting a learning goal (Ball, Thames, & Phelps, 2008).

To illustrate the unique measurement approach represented by the Learning Mathematics for Teaching (LMT) measures, an example item is presented in the Appendix. This item illustrates the difference between knowing simply the mathematics students are learning and knowing the specialized content needed to teach the mathematics to students. This distinction is important for considering the relevant assessment outcomes to use in studies of mathematics professional development. While the basic mathematics that students are expected to learn is clearly critical for teaching, this mathematics is typically not the primary focus of mathematics professional development. Instead, mathematics professional development is typically focused on the specialized types of mathematics encountered when teaching this content. As illustrated in the example in the Appendix, to provide a correct answer to this particular test item, teachers need to use their mathematical knowledge to select an appropriate story problem to illustrate a mathematics calculation; this is both different and more demanding than simply providing the correct answer to the story problem. Because MKT measures are broadly representative and closely tied to both the content of professional development and to teaching practice, they are well suited for assessing the outcomes of professional development and examining differences in knowledge that are associated with other valued outcomes (e.g., Hill, Rowan, & Ball, 2005).

The measures used in this study drew on computer implementations of the MKT assessments that allowed professional development providers to assess teachers' content knowledge for teaching. Several parallel forms were developed for each of the five different mathematics outcomes and were scored using item response theory. The reliabilities of these assessments ranged from 0.74 to 0.90 (see http://www.umich.edu/~lmtweb/ for more information) and forms were equated using item response linking methods (Hill & Ball, 2004; Hill, Ball, & Schilling, 2008; Hill, Schilling, & Ball, 2004). These assessments were delivered to participants through the online Teacher Knowledge Assessment System (TKAS) at the beginning and end of their professional development programs.

Our analyses in this study focused on knowledge in five different content areas that are common in mathematics professional development: (a) elementary school number concepts and operations;

(b) elementary school patterns, functions, and algebra; (c) elementary geometry; (d) middle school number concepts and operations; and (e) middle school patterns, functions, and algebra.

In addition to posttest scores from the MKT measure, the analyses examined several covariates collected at the onset of the professional development programs to understand how they might improve the sensitivity of designs in terms of the minimum detectable effect size they yield. We considered covariance adjustment for three different pretreatment covariates: (a) pretest, (b) teacher certification in mathematics education, and (c) teacher experience. We considered these particular covariates because prior studies have found them to be moderately predictive of teachers' knowledge.

For the pretest, teachers were measured in the same knowledge area as their post-test just before they participated in a professional development program. Teacher certification and experience were taken from a survey that accompanied the pretest. Certification was a simple indicator of whether a teacher has a permanent or professional certification in the area mathematics and was self-reported by teachers. We described teaching experience using 3 categories: up to 3 years of teaching, 4–15 years of teaching, or more than 15 years teaching.

Missing data

Given the large and diverse nature of our sample, the study sustained missing data within each sample originating from item nonresponse at the teacher level (e.g., Schafer & Graham, 2002). Across all samples, there was missing item information on approximately 13% of the variables (there was no unit nonresponse, only item nonresponse). Rather than remove cases with incomplete data, we used the three-level multiple imputation methods detailed in Asparouhov and Muthen (2010). We used an unrestricted multilevel imputation model that paralleled the three-level conditional analytic model detailed below (see equation 7) but included all of the aforementioned variables in our study. Our imputation approach implicitly assumes that data are missing at random (MAR), conditional upon the observed variables (Little & Rubin, 2002). As with all studies that draw upon multiple imputation and this assumption, because MAR is an untestable assumption, the validity of our subsequent results depends on the strength of this assumption under the observed variables. We then generated five separate imputed data sets, each containing different plausible values of the missing data points. Subsequent investigations repeated analyses using each of these five data sets. We present the pooled results across the five imputations and the corresponding uncertainty using appropriate procedures (Little & Rubin, 2002).

Analyses

Our investigation into design considerations for multilevel professional development studies was divided into three primary components. The first component assembled empirical estimates of the intraclass correlation coefficients and the second examined variance explained by covariates. The third component applied these estimates to the design of professional development studies and assessed the comparative advantage of several designs through an evaluation of the minimum detectable effect sizes.

Empirical estimates

To address the first component, we estimated intraclass correlation coefficients and the variance reduction power of covariates using a series of hierarchical linear models (Raudenbush & Bryk, 2002). For the two-level designs with teachers nested within schools (ignoring districts), we used the following unconditional specification:

$$\begin{aligned} Y_{ij}^{(A)} &= \pi_{0j} + \varepsilon_{ij} \qquad \varepsilon \sim N(0, \sigma_1^2) \\ \pi_{0j} &= \beta_{00} + u_{0j} \qquad u \sim N(0, \sigma_2^2), \end{aligned} \tag{1}$$

where $Y_{ij}^{(A)}$ is the post-program knowledge level for knowledge outcome A for teacher i in school j; π_{0j} is the average knowledge score for teachers in school j; ε_{ij} is the teacher specific error with an assumed

normal distribution centered at zero with variance σ_1^2; β_{00} is the grand mean knowledge score; and u_{0j} is the school specific random effect for school j with an assumed normal distribution with mean zero and variance σ_2^2. The unconditional intraclass correlation coefficient was estimated as

$$\rho_2 = \sigma_2^2 \,/\, (\sigma_1^2 + \sigma_2^2), \tag{2}$$

where ρ_2 is the school-level intraclass correlation coefficient and the denominator describes the total unconditional variation in the outcome. The variances of the residuals, σ_1^2, and school random effects, σ_2^2, also serve as the baseline estimates from which we will estimate the proportion of variation explained by covariates in the conditional models.

For the three-level designs with teachers nested within schools nested within districts, we used the following unconditional specification:

$$\begin{aligned} Y_{ijk}^{(A)} &= \pi_{0jk} + \varepsilon_{ijk} \qquad \varepsilon \sim N(0, \sigma_1^2) \\ \pi_{0jk} &= \beta_{00k} + u_{0jk} \qquad u \sim N(0, \sigma_2^2) \\ \beta_{00k} &= \gamma_{000} + r_{00k} \qquad r \sim N(0, \sigma_3^2), \end{aligned} \tag{3}$$

where $Y_{ijk}^{(A)}$ is the knowledge level for knowledge outcome A for teacher i in school j in district k; π_{0jk} is the average knowledge score for teachers in school j in district k; ε_{ijk} is the teacher-specific error; β_{00k} is the average score in school k; u_{0jk} is the school-specific random effect for school j; γ_{000} is the overall average knowledge score; and r_{00k} is district-specific random effect for district k with variance σ_3^2. We now have two unconditional intraclass correlation coefficients:

$$\begin{aligned} \rho_2 &= \sigma_2^2/(\sigma_1^2 + \sigma_2^2 + \sigma_3^2) \\ \rho_3 &= \sigma_3^2/(\sigma_1^2 + \sigma_2^2 + \sigma_3^2), \end{aligned} \tag{4}$$

where ρ_2 is the proportion of outcome variance attributable to differences among schools and ρ_3 is the proportion of outcome variance attributable to differences among districts. Similar to the two-level model, these variance estimates serve as the baseline estimates from which we estimated the proportion of variation explained by covariates in the conditional models.

Because our estimates are based on varying sample sizes and small changes in the intraclass correlation coefficient can impact the efficiency of designs, we also estimated confidence intervals using a multilevel parametric bootstrap. We describe the uncertainty of the intraclass correlation coefficient estimates by constructing empirical confidence intervals (95%) using exact percentiles (i.e., 2.5% and 97.5%) of 1,000 bootstrapped estimates.

To assess the variance explained by several pretreatment covariates, we estimated the change in variance due to covariance adjustment. We modified the unconditional models to include teacher covariates and their aggregates at higher levels. For two-level models we used

$$\begin{aligned} Y_{ij}^{(A)} &= \pi_{0j} + \pi_1(X_{ij} - \overline{X}_j) + \varepsilon_{ij} \qquad \varepsilon \sim N(0, \sigma_{\tilde{1}}^2) \\ \pi_{0j} &= \beta_{00} + \beta_1 \overline{X}_j + u_{0j} \qquad u \sim N(0,\ \sigma_{\tilde{2}}^2). \end{aligned} \tag{5}$$

Continuing with the aforementioned notation, we used X_{ij} as a covariate for teacher i in school j and $\overline{X}_j$ as the school-specific mean of this covariate. In this model, our estimates of the residual variance, $\sigma_{\tilde{1}}^2$, and the variance of the school random effects, $\sigma_{\tilde{2}}^2$, represent the adjusted variation in the outcome conditional upon the respective covariates. The reduction in variance supplied by the covariate(s) using

the respective variance components estimated is

$$R^2_{\cdot} = (\sigma^2_{\cdot} - \sigma^2_{\tilde{\cdot}})/\sigma^2_{\cdot}, \tag{6}$$

where $R^2_{\cdot}$ represents the reduction in teacher- (R^2_1) or school-level (R^2_2) outcome variance attributable to the covariates, respectively.[1]

For three-level models we used

$$\begin{aligned} Y^{(A)}_{ijk} &= \pi_{0jk} + \pi_1(X_{ijk} - \overline{X}_{jk}) + \varepsilon_{ijk} \qquad \varepsilon \sim N(0, \sigma^2_{\tilde{1}}) \\ \pi_{0jk} &= \beta_{00k} + \beta_1(\overline{X}_{jk} - \overline{X}_k) + u_{0jk} \qquad u \sim N(0,\ \sigma^2_{\tilde{2}}) \\ \beta_{00k} &= \gamma_{000} + \gamma_{001}\overline{X}_k + r_{00k} \qquad r \sim N(0, \sigma^2_{\tilde{3}}), \end{aligned} \tag{7}$$

We now introduce $\sigma^2_{\tilde{3}}$ as the adjusted district-level variation and we used X_{ij} as a covariate for teacher i in school j in district k, with $\overline{X}_{jk}$as its school-specific mean and $\overline{X}_{jk}$ as its district-specific mean. The reduction in variance attributable to a covariate at each level was estimated using formulas similar to expression (6).

Design implications

In the second component of our analysis we explored how we might use the aforementioned results to assess and improve the sensitivity of designs. More specifically, a common goal of study design is to identify the design most sensitive to potential differences between the treatment groups given fixed constraints. Although there are multiple approaches to describe the sensitivity of a design, we summarized the sensitivity of designs using minimum detectable effect sizes (see details below).

We described design sensitivity through the minimum detectable effect size for two main reasons. First, the use of the minimum detectable effect size metric is widespread and has a direct correspondence to alternative design sensitivity measures such as power (Bloom, 2008). Second, the minimum detectable effect size can be especially useful for outcomes that have little or no empirical basis for interpreting effect size. More specifically, there is a growing body of research that has argued against effect size rules of thumb in favor of anchoring effect sizes in empirical estimates specific to the context under study (e.g., Hill, Bloom, Black, & Lipsey, 2008). Because there is a lack of evidence describing the empirical distribution of effect sizes in professional development, we report minimum detectable effect sizes under different sample sizes and cost structures rather than select arbitrary effect sizes and report, for example, power.

The analyses below demonstrate how to use the estimates of the intraclass correlation coefficients and variance explained by a covariate to evaluate minimum detectable effect sizes and assess the relative advantage of covariance adjustment on different covariates under different cost structures. We outline the designs and analytic approaches we considered in assessing the sensitivity of designs and their relative rank order for a given set of constraints.

Designs considered

We sampled five types of designs[2]: (a) two-level teacher-randomized designs that block on schools and allow for effects to vary across schools but ignored districts, (b) two-level school-randomized designs that ignored districts, (c) three-level district-randomized designs, (d) three-level school-randomized designs that block on districts and allow effects to vary across districts, and (e) three-level school-randomized designs that block on districts and constrain the variability in treatment effects to zero. The

[1] R^2 values are not well defined in multilevel models because variance can increase with the introduction of covariates. Our approach follows the literature by using pseudo-R^2 values to describe the reduction in variation (Hedges & Hedberg, 2007; Jacob, Zhu, & Bloom, 2010; Spybrook & Raudenbush, 2009).

[2] For a point of reference, we also considered teacher-randomized designs that ignore schools and districts.

purpose in using these designs was to illustrate the use of the parameter estimates in several common designs that lend themselves to the multilevel nature of professional development rather than to comprehensively examine all possible designs and scenarios. By contrasting these designs, we intended to demonstrate the conditions under which covariance adjustment and blocking on school or district membership might improve the absolute and relative sensitivity of a design. We describe the designs in more detail below.

We adopted the following specification for two-level teacher-randomized designs that block on schools:

$$\begin{aligned} Y_{ij}^{(A)} &= \pi_{0j} + \delta_{0j}T_{ij} + \pi_1(X_{ij} - \overline{X}_j) + \varepsilon_{ij} && \varepsilon \sim N(0, \sigma^2_{1_\delta}) \\ \pi_{0j} &= \beta_{00} + \beta_1\overline{X}_j + u_{0j} && u_{0j} \sim N(0, \sigma^2_{2_\delta}) \\ \delta_{0j} &= \delta + u_{1j} && u_{1j} \sim N(0, \tau_2^2) \end{aligned} \tag{8}$$

Continuing with the aforementioned notation we now introduce T_{ij} as the treatment indicator and δ as the estimated overall treatment effect and u_{1j} as the school-specific change in the overall treatment effect with variance τ_2^2. We use $\sigma^2_{1_\delta}$ and $\sigma^2_{2_\delta}$ to denote the levels one and two variance components for the outcome when they are conditional upon the treatment and covariates and use $\sigma^2_{1_\delta}$ and $\sigma^2_{2_\delta}$ when the variance components are conditional upon only the treatment.

Similarly, using the same notation we specified the two-level school-randomized experiments with teachers nested within schools (ignoring districts) as:

$$\begin{aligned} Y_{ij}^{(A)} &= \pi_{0j} + \pi_1(X_{ij} - \overline{X}_j) + \varepsilon_{ij} && \varepsilon \sim N(0, \sigma^2_{1_\delta}) \\ \pi_{0j} &= \beta_{00} + \delta T_j + \beta_1\overline{X}_j + u_{0j} && u \sim N(0, \sigma^2_{2_\delta}). \end{aligned} \tag{9}$$

For three-level district randomized designs that assume teachers are nested within schools nested within districts, we used the following specification:

$$\begin{aligned} Y_{ijk}^{(A)} &= \pi_{0jk} + \pi_1(X_{ijk} - \overline{X}_{jk}) + \varepsilon_{ijk} && \varepsilon \sim N(0, \sigma^2_{1_\delta}) \\ \pi_{0jk} &= \beta_{00k} + \beta_1(\overline{X}_{jk} - \overline{X}_k) + u_{0jk} && u \sim N(0,\ \sigma^2_{2_\delta}) \\ \beta_{00k} &= \gamma_{000} + \delta T_k + \gamma_{001}\overline{X}_k + r_{00k} && r \sim N(0, \sigma^2_{3_\delta}), \end{aligned} \tag{10}$$

using notation following that of earlier models.

For three-level school randomized designs that block on district and allow treatment effect variability via random district effects, we used the following multidistrict school-randomized model:

$$\begin{aligned} Y_{ijk}^{(A)} &= \pi_{0jk} + \pi_1(X_{ijk} - \overline{X}_{jk}) + \varepsilon_{ijk} && \varepsilon \sim N(0, \sigma^2_{1_\delta}) \\ \pi_{0jk} &= \beta_{00k} + \delta_{00k}T_{jk} + \beta_1(\overline{X}_{jk} - \overline{X}_k) + u_{0jk} && u \sim N(0, \sigma^2_{2_\delta}) \\ \beta_{00k} &= \gamma_{000} + \gamma_{001}\overline{X}_k + r_{00k} && r_{00k} \sim N(0, \sigma^2_{3_\delta}) \\ \delta_{00k} &= \delta + r_{01k} && r_{01k} \sim N(0, \tau_3^2)\ , \end{aligned} \tag{11}$$

with r_{01k} representing the district-specific treatment effect deviation from the overall average treatment effect, δ, and τ_3^2summarizing the variability of these deviations. For the final design (E) that constrained effect variation to zero, we drew on fixed district effects. Specifically, we amended the previous expression (11) so that r_{00k} represented fixed effects (and thus dropping $\overline{X}_k$) and constraining r_{01k} to zero (no treatment variation across districts).

Optimal sample allocation

A major factor driving the sensitivity of a design is the sample size. Although the sample sizes at each level contribute to the sensitivity of a design, the relative contribution of the sample sizes is governed, in part, by the variance decomposition and the relative cost of sampling units at each level. Literature has developed approaches to identify the most efficient (smallest error variance) or optimal sampling strategy given the intraclass correlation coefficient(s) and the costs of sampling units at each level (e.g., Raudenbush, 1997).

To identify the sampling strategies that produce the smallest error variance for a fixed set of intraclass correlation coefficients and cost structure, we adopted a framework such that the total cost of collecting data for a study was a linear function of the sample size at each level (Raudenbush, 1997). Using c_1 as the cost per teacher, c_2 as the cost per school, c_3 as the cost per district, n_1 as the number of teachers per school, n_2 as the number of schools per district, and n_3 as the number of districts, the total cost of the study is summarized by (for designs that ignore districts we assume districts do not contribute to the total cost)

$$C = c_3 n_3 + c_2 n_3 n_2 + c_1 n_3 n_2 n_1 \tag{12}$$

The sample allocations across levels that produce the minimum variance of the treatment estimate are as follows (Hedges & Borenstein, 2014). For two-level teacher-randomized within blocks defined by the schools design, the optimal number of teachers per school to sample, n_1^O, is

$$n_1^O = 2\sqrt{\frac{c_2}{2c_1}\left(\frac{(1-R_1^2)(1-\rho_{2_\delta})}{(1-R_T^2)\omega_2\rho_{2_\delta}}\right)}, \tag{13}$$

where $\rho_{2_\delta} = \frac{\sigma^2_{1_\delta}}{\sigma^2_{1_\delta}+\sigma^2_{2_\delta}}$, $\omega_2 = \frac{\tau_2^2}{\sigma^2_{2_\delta}}$, and $R^2_{T_2} = \frac{\tau_2^2 - \tilde{\tau}_2^2}{\tau_2^2}$ so that τ_2^2 is the unconditional treatment effect variation across sites and $\tilde{\tau}_2^2$ is the treatment effect variation across sites conditional on the covariates. The resulting two-level sample size for a given budget can then be recovered using the total cost expression. For two-level school-randomized designs that ignore districts, the optimal number of teachers per school to sample, n_1^O, is

$$n_1^O = \sqrt{\frac{c_2}{c_1}\left(\frac{(1-R_1^2)(1-\rho_{2_\delta})}{(1-R_2^2)\rho_{2_\delta}}\right)} \text{ where } \rho_{2_\delta} = \frac{\sigma^2_{1_\delta}}{\sigma^2_{1_\delta}+\sigma^2_{2_\delta}}. \tag{14}$$

For three-level district-randomized designs, the optimal number of teachers per school, n_1^O, and the optimal number of schools per district, n_2^O, is

$$n_1^O = \sqrt{\frac{c_2}{c_1}\left(\frac{(1-R_1^2)(1-\rho_{2_\delta}-\rho_{3_\delta})}{(1-R_2^2)\rho_{2_\delta}}\right)} \text{ and } n_2^O = \sqrt{\frac{c_3}{c_2}\left(\frac{(1-R_2^2)(\rho_{2_\delta})}{(1-R_3^2)\rho_{3_\delta}}\right)}, \tag{15ab}$$

where $\rho_{2_\delta} = \frac{\sigma^2_{2_\delta}}{\sigma^2_{1_\delta}+\sigma^2_{2_\delta}+\sigma^2_{3_\delta}}$ and $\rho_{3_\delta} = \frac{\sigma^2_{3_\delta}}{\sigma^2_{1_\delta}+\sigma^2_{2_\delta}+\sigma^2_{3_\delta}}$. Under three-level school-randomized experiments

that block on districts, the optimal sample schemes are

$$n_1^O = \sqrt{\frac{c_2}{c_1}\left(\frac{(1-R_1^2)(1-\rho_{2_\delta}-\rho_{3_\delta})}{(1-R_2^2)\rho_{2_\delta}}\right)} \text{ and } n_2^O = 2\sqrt{\frac{c_3}{2c_2}\left(\frac{(1-R_2^2)\rho_2}{(1-R_{T_3}^2)\omega_3\rho_3}\right)}, \tag{16ab}$$

where τ_3^2 is the unconditional treatment effect variation across sites and $\tilde{\tau}_2^2$ is the treatment effect variation across sites conditional on the covariates such that $R_{T_3}^2 = \frac{\tau_3^2 - \tilde{\tau}_3^2}{\tau_3^2}$ and $\omega_3 = \frac{\tau_3^2}{\sigma_{3_\delta}^2}$.

Covariate cost threshold

We next explored the relative sensitivity of the proposed designs by considering differential costs per unit under unconditional and conditional designs. We developed a type of cost inflation threshold that captures the point at which the cost of collecting a covariate overtakes the reduction in variance it supplies such that a conditional design results in a net increase in minimum detectable effect size.

Literature has generally supported designs that measure and adjust for pretreatment covariates because they tend to increase efficiency and decrease the minimum detectable effect size (e.g., Raudenbush, Martinez, & Spybrook, 2007). However, literature has not considered the potential cost tradeoffs associated with collecting pre-treatment covariates. For instance, although designs that adjust for a pre-test may be more efficient than unconditional designs, it is likely that designing, collecting, and/or administering a pre-test may increase the cost per unit. Under a fixed budget, these increased costs may result in a smaller sample size for the conditional design that includes a teacher-level pre-test than for the unconditional design and, ultimately, yield a less sensitive design compared to unconditional designs.

Our examination begins by continuing with the conventional optimal design framework outlined above (e.g., Raudenbush, 1997). We then simply modify this cost structure to incorporate the potential for differential per unit data collection costs under designs that condition on no covariates or different covariates. In turn, we use the adjustments to explore how we might assess the increased costs associated with collecting covariate information under three simple scenarios.

The first scenario considers a simple increase in the cost of sampling teachers (only). For instance, this cost structure might be appropriate when additional costs per teacher are incurred because of collecting teacher questionnaire information from each teacher. We modified the aforementioned cost structure so that the cost per unit was more for designs that collected a covariate (e.g., pre-test, teacher certification) than for an unconditional specification. The cost of sampling a teacher under an unconditional specification was c_1 but the cost of sampling a teacher when collecting a covariate was k_1c_1. Here we use k_1 as a cost inflation factor. For example, when k_1 equals 2, the cost per teacher under a conditional specification (i.e., collecting a covariate) would be twice as much as that of an unconditional specification (i.e., not collecting covariate information). Under this situation, the total cost of a study with a covariate would be

$$C = c_3n_3 + c_2n_3n_2 + k_1c_1n_3n_2n_1. \tag{17}$$

In turn, we defined the covariate cost inflation threshold as the inflation factor for which the minimum detectable effect size for the covariate design is equal to that of the unconditional design.

In the second scenario, we modified the cost structure such that the cost per school was more for designs that collected covariates than for an unconditional design. In contrast to the first scenario, this cost structure might arise because the cost of administering a professional development survey introduces fixed additional costs at the school level regardless of how many teachers in that school participate (e.g., costs associated with a school-wide survey administrator). The cost of sampling a school under an unconditional design was c_2 and the cost of sampling each school when collecting a covariate was k_2c_2, where k_2 was a cost inflation factor. Under this situation, the total cost of a study with a covariate

would be

$$C = c_3 n_3 + k_2 c_2 n_3 n_2 + c_1 n_3 n_2 n_1. \tag{18}$$

For the third scenario, we retained the above cost structures but considered the relative efficiency of collecting two different covariates. Specifically, we explored when and to what extent designs that use covariance adjustment on an economical covariate with lower explanatory power outperform designs that use a more costly covariate with higher explanatory power. For instance, this scenario might arise when we have a choice between collecting a covariate that can be measured easily through a survey (e.g., teacher certification status) and collecting a covariate that takes considerably more effort and cost (e.g., pre-test) but is likely to have a higher correlation with the outcome.

To identify covariate cost inflation thresholds under this example, we assumed the default cost structure (expression 12) applied to the economical covariate with lower explanatory power (e.g., teacher certification status) and the modified cost structure (expression 17 or 18) applied to the more costly covariate with higher explanatory power (e.g., pre-test). Like the previous scenarios, we defined the covariate cost inflation threshold as the inflation factor for which the minimum detectable effect size for the economical covariate design is equal to that of the costly covariate design.

Minimum detectable effect size. As noted earlier, our analyses focused on the minimum detectable effect sizes to assess the sensitivity of designs (Bloom, 2008). The minimum detectable effect size describes the smallest possible effect size a design can detect for a given power level and sample size. To assess the absolute and relative sensitivity of the designs in terms of their minimum detectable effect sizes, we considered balanced designs with 80% power under a two-tailed test with a type one error rate of 0.05 (e.g., Hedges & Hedberg, 2007).

For two-level teacher-randomized designs that block on schools and use random school effects, we estimated minimum detectable effect sizes (MDES) using

$$MDES = M_{(n_2 - 1)} * \sqrt{\frac{\omega_2 \rho_2}{n_2} + \frac{(1-\rho_2)(1-R_1^2)}{P(1-P)n_1 n_2}}, \tag{19}$$

where n_2 is the number of schools, n_1 is the number of teachers within a school, M is a design multiplier such that $M \approx t_{\alpha/2} + t_{1-\beta}$, where t is the critical value and α and β are the type one and two error rates, for a t distribution with n_2-1 degrees of freedom, P is the proportion of schools assigned to treatment, τ_2^2 is variance among the school-specific deviations in the overall treatment effect, ρ_2 is the intraclass correlation coefficient, and R_1^2 is the teacher variance explained.

Under the two-level school-randomized design with teachers nested within schools, we estimated minimum detectable effect sizes using

$$MDES = \frac{M_{(n_2 - 2 - p_2)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\rho_2(1-R_2^2)}{n_2} + \frac{(1-\rho_2)(1-R_1^2)}{n_1 n_2}}, \tag{20}$$

with p_2 as the number of level two covariates and R_2^2 as the variance explained at level two.

Under the three-level district-randomized design with teachers nested within schools nested within districts, we estimated minimum detectable effect sizes using

$$MDES = \frac{M_{(n_3 - 2 - p_3)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\rho_3(1-R_3^2)}{n_3} + \frac{\rho_2(1-R_2^2)}{n_2 n_3} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{n_1 n_2 n_3}}, \tag{21}$$

where n_3 is the number of districts, n_2 is the number of schools within a district, n_1 is the number of

teachers within a school, ρ_2 and ρ_3 are the proportions of total variance attributable to levels two and three, respectively, and R_3^2 as the variance explained at the district level. Remaining notation follows from previous formulas.

For three-level school-randomized designs that block on district and use random district effects, we estimated minimum detectable effect sizes using (with similar notation).

$$MDES = M_{(n_3-1)} * \sqrt{\frac{\omega_3\rho_3}{n_3} + \frac{\rho_2(1-R_2^2)}{P(1-P)n_2n_3} + \frac{(1-\rho_2-\rho_3)(1-R_1^2)}{P(1-P)n_1n_2n_3}}. \quad (22)$$

Results

The results are divided into two sections that align with the research questions and methods. In the first section, we summarized the empirical estimates of the design parameters from our sample for each outcome using the two- and three-level models. The second section then surveys the use of these estimates to inform and improve the design of professional development studies.

Empirical estimates

The variance decompositions for the two- and three-level models suggested substantial clustering among teachers both within schools and within districts (Table 2). Two-level variance decompositions using teachers nested within schools (ignoring districts) 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 2). Three-level variance decompositions with teachers nested within schools nested within districts indicated that the variance attributable to (a) 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) 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).

Table 2. Two-level (teachers nested within schools) and three-level (teachers nested within schools nested within districts) unconditional intraclass correlation coefficients (ICC) and their confidence intervals by outcome.

Outcome	ICC	Low	High
Two-level (schools)			
Grades 4–8 geometry	.33	.27	.37
Elementary school number concepts and operations	.19	.16	.21
Elementary school patterns, functions, and algebra	.29	.25	.32
Middle school number concepts and operations	.17	.11	.22
Middle school patterns, functions, and algebra	.27	.24	.30
Three-level (schools)			
Grades 4–8 geometry	.21	.15	.25
Elementary school number concepts and operations	.03	.01	.05
Elementary school patterns, functions, and algebra	.05	.02	.07
Middle school number concepts and operations	.03	.00	.09
Middle school patterns, functions, and algebra	.14	.10	.17
Three-level (districts)			
Grades 4–8 geometry	.11	.03	.17
elementary school number concepts and operations	.13	.10	.16
Elementary school patterns, functions, and algebra	.18	.13	.22
Middle school number concepts and operations	.13	.07	.17
Middle school patterns, functions, and algebra	.13	.09	.16

Notes. ICC is intraclass correlation coefficient. 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 covariate for each outcome by level for two- and three-level specifications.

	Pre-test			Certification			Experience		
Outcomes	Teacher	School	District	Teacher	School	District	Teacher	School	District
Two-level									
EG	.09	.25	—	.00	.19	—	.00	.00	—
ESNCO	.05	.21	—	.00	.04	—	.00	.00	—
ESPFA	.03	.14	—	.00	.05	—	.00	.01	—
MSNCO	.09	.24	—	.02	.01	—	.00	.00	—
MSPFA	.06	.36	—	.01	.10	—	.01	.01	—
Three-level									
EG	.09	.32	.08	.00	.08	.38	.00	.01	.00
ESNCO	.05	.21	.16	.00	.01	.04	.00	.00	.00
ESPFA	.03	.11	.10	.00	.02	.07	.00	.00	.03
MSNCO	.09	.60	.22	.02	.00	.06	.00	.26	.01
MSPFA	.06	.46	.23	.01	.20	.00	.01	.00	.03

Notes. ESNCO: elementary school number concepts and operations; ESPFA: elementary school patterns, functions, and algebra; EG: elementary geometry; MSNCO: middle school number concepts and operations; and MSPFA: middle school patterns, functions, and algebra. R^2 is the proportion of variation explained.

Subsequent analyses suggested that a moderate proportion of the variance was accounted for by teachers' prior abilities and, to a lesser extent, teacher certification and teachers' experience (Table 3). More specifically, the prognostic value of the pre-test was generally small at the teacher and district levels and moderate at the school level. On average the correlation between the pre- and post-test at the two level was 0.5. However, this correlation varied considerably across outcomes. For instance, although prior ability accounted for as much as 60% of the school-level variance in middle school number concepts and operations, it accounted for just over 20% of the school-level variance in elementary school number concepts and operations. The explanatory value of teacher certification status was also inconsistent across levels and outcomes but explained about two-thirds less variance than the pre-test. Similarly, teacher experience explained virtually no variance.

Assessing design implications

Having assembled initial empirical estimates of the parameters needed to design multilevel studies of professional development, we next illustrate how to use these estimates to evaluate and compare the sensitivity of the proposed designs under particular conditions. Although each of the designs considered draws on the multilevel structure of schooling as a primary design consideration, the scope of inference varies across designs. For instance, multisite or blocked designs incorporate consideration of treatment effect variability across hierarchical units whereas this parameter is absent from cluster-randomized designs. Because our investigation of relative sensitivity involves designs with different scopes of inference, the sampling scheme that produces the most efficient (minimum error variance) estimate may differ across designs. To facilitate comparisons across designs, we used the minimum detectable effect sizes under the optimal sample allocations particular to each design.

To make the comparisons in our illustration more concrete, we adopted a fixed cost structure that was similar to structures previously published in the literature concerning the design of multilevel studies (Hedges & Borenstein, 2014; Raudenbush, 1997; Raudenbush & Liu, 2000). In particular, we assumed the following cost structure: total cost available for study (C) = 1000, cost per teacher ($c1$) = 1, cost per school ($c2$) = $5c1$, and cost per district ($c3$) = $5c2$ and (when relevant) the ratio of variance of the treatment effect across sites to the variance of the site means (ω) was 0.05. Using the estimated intraclass correlation coefficients and the implied optimal sample sizes, the corresponding minimum detectable effect sizes for unconditional designs are presented in Table 4. We again note that the results are illustrative and specific to our assumed cost structure and parameter values and do not necessarily generalize to alternative assumptions. Adjustments for different cost structures should be made in investigating the tradeoffs associated with a particular study.

Table 4. Minimum detectable effect sizes for five unconditional designs using the optimal sample sizes for the example cost structure.

		Sample Sizes by Design					Minimum Detectable Effect Sizes by Design				
Outcome	Unit	A	E	B	D	C	A	E	B	D	C
Grade 4–8 geometry	Teachers	20	4	3	4	4	0.21	0.38	0.38	0.63	0.66
	Schools	40	20	122	20	3					
	Districts	—	5	—	5	19					
Elementary school number concepts and operations	Teachers	29	12	5	12	12	0.22	0.28	0.34	0.44	0.58
	Schools	29	7	104	7	1					
	Districts	—	7	—	7	23					
Elementary school patterns, functions and algebra	Teachers	22	9	3	9	9	0.21	0.31	0.37	0.45	0.65
	Schools	37	7	118	7	1					
	Districts	—	8	—	8	24					
Middle school number concepts and operations	Teachers	31	12	5	12	12	0.23	0.28	0.33	0.44	0.58
	Schools	28	7	101	7	1					
	Districts	—	7	—	7	23					
Middle school patterns, functions and algebra	Teachers	23	5	4	5	5	0.21	0.35	0.36	0.56	0.65
	Schools	35	15	115	15	2					
	Districts	—	6	—	6	21					

Notes. The minimum detectable effect size under the teacher-randomized design that ignores school and districts using the sample sizes associated with design A was between 0.18 and 0.19, depending on outcome. Cost structure: $C = 1000$, $c1 = 1$, $c2 = 5c1$, $c3 = 5c2$.$\omega = 0.05$.
A: two-level teacher-randomized designs that block on schools
B: two-level school-randomized designs
C: three-level district-randomized designs
D: three-level school-randomized designs that block on districts (random district effects)
E: three-level school-randomized designs that block on districts (fixed district effects)

In contrasting the minimum detectable effect sizes across designs, the results demonstrated consistent and substantial differences across all outcomes (Table 4). Perhaps the most notable result was the well-established pattern of the lower the unit of assignment, the more sensitive the design. This result was expected (Raudenbush, 1997). For instance, designs that assigned treatments to teachers had smaller minimum detectable effect sizes than designs that assigned treatments to schools or districts. Across all outcomes, the average minimum detectable effect sizes (in rank order) were two-level teacher-randomized within school blocks (A): 0.22; the three-level school-randomized within district blocks design assuming no effect variability (B): 0.32; the two-level school-randomized design (C): 0.36; the three-level school-randomized within district blocks design assuming effect variability (D): 0.50; and the three-level district-randomized design (E): 0.62.

Next, we examined the absolute and relative sensitivity of the designs under covariance adjustment. Table 5 presents the minimum detectable effect size results but omits the optimal sample sizes because

Table 5. Minimum detectable effect sizes for five conditional designs using the optimal sample sizes for the example cost structure.

	Pre-test					Certificate				
Outcome	A	E	B	D	C	A	E	B	D	C
Grades 4–8 Geometry	0.20	0.33	0.34	0.60	0.61	0.21	0.37	0.35	0.69	0.59
Elementary school number concepts and operations	0.22	0.27	0.31	0.44	0.54	0.22	0.28	0.33	0.46	0.57
Elementary school patterns, functions, and algebra	0.21	0.30	0.35	0.46	0.62	0.21	0.31	0.36	0.47	0.64
Middle school number concepts and operations	0.22	0.25	0.30	0.40	0.50	0.22	0.28	0.33	0.46	0.57
Middle school patterns, functions, and algebra	0.21	0.30	0.31	0.51	0.57	0.21	0.33	0.35	0.57	0.64

Notes. Cost structure: $C = 1000$, $c1 = 1$, $c2 = 5c1$, $c3 = 5c2$.$\omega = 0.05$.
A: two-level teacher-randomized designs that block on schools
B: two-level school-randomized designs
C: three-level district-randomized designs
D: three-level school-randomized designs that block on districts (random district effects)
E: three-level school-randomized designs that block on districts (fixed district effects)

they were very similar to the unconditional optimal sample sizes (Table 4). The minimum detectable effect size associated with designs that use covariance adjustment preserved the rank ordering of the unconditional designs. The advantage of adjusting on the pre-test or certification varied by design; however, overall the value was modest. On average, adjusting for the pre-test lowered the minimum detectable effect size between 1% and 10%.

Finally, we investigated the relative sensitivity of the designs to additional covariate costs through the covariate cost inflation threshold strategy described above. Results across the three aforementioned scenarios and five outcomes were qualitatively similar. As a result, we limit our illustration to the first scenario (increased teacher costs) using the geometry outcome. As in previous examples, because the results are subject to the assumed cost structure and parameter estimates, the example is meant to provide an illustration of the proposed strategy and should not be generalized to alternative sets of assumptions.

We continued with the previous cost structure but assumed that the increased cost for collecting a covariate only impacted the cost of sampling teachers. We present the results under the empirical estimates of the intraclass correlation coefficients and variance explained and the corresponding optimal sample sizes in Figure 1. For each design, we plotted the minimum detectable effect sizes for unconditional, certification adjusted, and pre-test-adjusted specifications.

Across all designs, we saw that the covariate cost-inflation-factor threshold ranged from about 1.1 to 2 (three-level district-randomized design) with an average of about 1.5; that is, the net improvement in the minimum detectable effect size offered by a pre-test or certification covariate was usually overtaken by the cost of collecting the covariates once the per-teacher cost increased by about 50%. For instance, under the two-level school-randomized design (B), the covariate cost inflation threshold was approximately 1.4 and 1.7 for the certification and pre-test specifications. Put differently, if collecting a certification/pre-test covariate increased per-teacher costs by 40% to 70%, the unconditional design is likely to be more sensitive given a fixed budget (Figure 1).

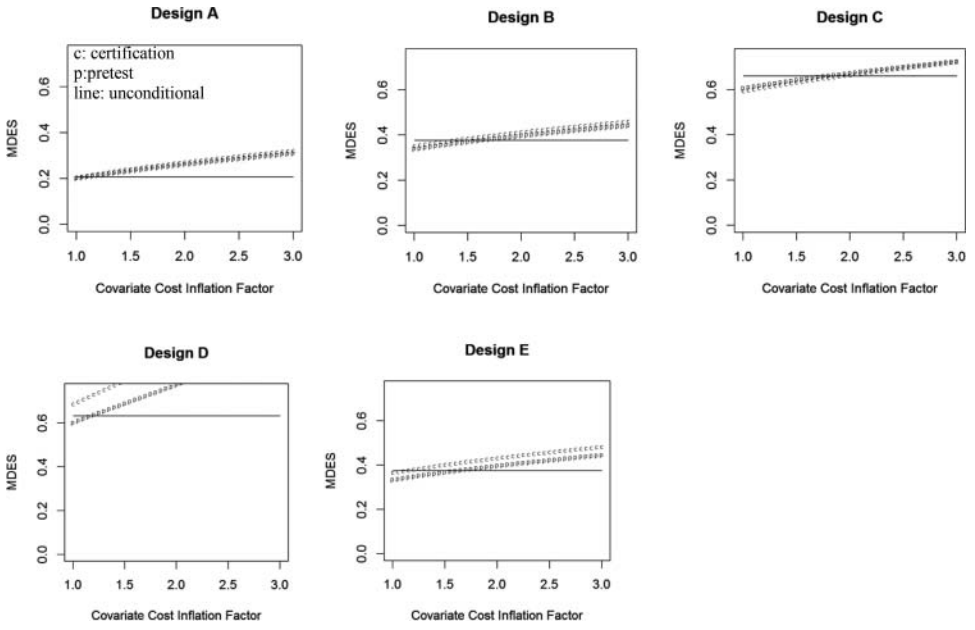

Figure 1. Minimum detectable effect size as a function of teacher-level covariate cost inflation factor for grades 4–8 geometry (relative to unconditional design).
Note. A: two-level teacher-randomized designs that block on schools.
B: two-level school-randomized designs.
C: three-level district-randomized designs.
D: three-level school-randomized designs that block on districts (random district effects).
E: three-level school-randomized designs that block on districts (fixed district effects).

Discussion

A critical consideration in the design of contemporary professional development studies is clarifying the types of designs that are amenable to the collaborative structure of many programs in order to provide a careful mapping of theories of action onto study designs. The objective of this study was to explore the alignment of several cluster-based designs and investigate the feasibility of these designs for studying professional development programs. In what follows, we synthesize the implications and issues that arise from the results, discuss limitations, and conclude with preliminary recommendations.

Implications

Broadly, the results have potentially important implications for designing studies of professional development. Perhaps the most prominent implication concerns the likely scale of data collection necessary to yield well-powered designs. For instance, to reach conventional power levels (i.e., 0.80), clustered designs will often need almost double the number of schools as compared to teacher-randomized designs. The collective results suggested that clustered designs will typically necessitate large sample sizes or anticipate large effects in order to provide sufficiently precise estimates of the effects. This finding has significant implications for researchers but also for the larger teacher professional development field because it suggests that experimental professional development studies will require more resources than are commonly available in current funding streams.

More specifically, typical educational experiments with student outcomes tend to sample about 50 schools (Schochet, 2011). In many cases, this school-level sample size provides a sufficient level of precision because, in part, researchers leverage two important design features. The first is conditioning on covariates that are highly predictive of student outcomes. The results of our study suggested that this strategy may be more limited in professional development studies because even the typically most predictive covariate, the pre-test, demonstrated fairly modest correlations with the outcome. Second, the accompanying student-level sample sizes in typical educational experiments are usually large enough (e.g., 20) to substantially reduce the residual contribution of the student-level error variance (i.e., the student-level variance component that adds to the uncertainty in the cluster means, σ_1^2/n_1) to the standard error of the estimated treatment effect. In contrast, because teacher-level sample sizes per school in professional development studies are likely to be significantly smaller, these same designs (i.e., 50 schools) will typically be underpowered.

Several practical design strategies are implicated by the necessary scale of these studies. One route is to find alternative prognostic covariates. However, research has found it difficult to predict teacher knowledge and effectiveness and so even this route may prove to be limited (e.g., MET, 2012). Another route is for the field to reserve clustered experimental designs for only those professional development programs that have shown significant promise, extensive field testing, and relatively homogeneous effects across clusters. Alternatively, clustered experimental designs may also be earmarked for only those programs that have demonstrated large effect sizes.

Lastly, the results suggest that one ostensible route to designing a well-powered experimental assessment of professional development programs is to use teacher-randomized designs. However, although the sensitivity of a design to detect an effect is an important consideration, in many instances the collaborative nature of a professional development program will supplant this consideration. In this way, the preference for teacher-randomized designs is more qualified. If the program under study excludes collaboration, the most sensitive design is likely to be a teacher-randomized design. However, if the program includes collaboration as a major component, it is likely that teacher-randomized designs will be unsuitable on theoretical grounds.

More technically, the results within and across designs and outcomes suggested that the unit of randomization played a much larger role in determining design sensitivity than covariance adjustment or blocking. For instance, the minimum detectable effect size for the district-randomized design (C) was typically three times larger than that of the teacher-randomized design (A). By comparison, the unconditional specifications under those designs presented minimum detectable effect sizes that were only

about 2%–9% larger than those under the covariance adjustment on the pre-test specification. This result reiterates the well-known result that cluster-randomized designs lack statistical precision when compared to individually randomized designs. At the same time, the results quantify the general magnitude of the loss of precision originating from the clustering, prognostic value of covariates, and teacher-level sample sizes.

Blocking played a less influential but more complex and two-dimensional role in shaping the sensitivity of a design. Blocking on schools/districts was a moderately effective strategy for improving design sensitivity, but its contribution was substantially moderated by assumptions concerning effect variability. More specifically, the variance eliminated by blocking across outcomes was moderate, averaging 0.25 for the unconditional teacher-randomized within school blocks design (A) and 0.14 for the school-randomized within district blocks designs (D and E). Ostensibly, this reduction in outcome variance should improve the overall sensitivity of a design. However, as previously described, blocked designs potentially introduce additional variance arising from treatment effect variability. The net result is that the reduction in outcome variance supplied by blocking can be offset by variation in the effect.

The impact of the dual and contrasting contributions of blocking to design sensitivity is apparent in our worked example (see Table 4). Take for instance the minimum detectable effect sizes of the school-randomized design (B), which ignores districts, and the school-randomized within district blocks designs (D and E). Across all outcomes, the average minimum detectable effect size of the school-randomized design (B) was 0.36. Advancing this design by blocking on districts and assuming no effect size variability (design E), the average minimum detectable effect size fell to 0.32. By contrast, blocking on districts but allowing effects to differ across districts (design D) raised the minimum detectable effect size to 0.50. In this way, the expanded inference space associated with blocking introduced important tradeoffs concerning the scope of a research study, the intended inference, and design sensitivity.

Covariance adjustment on the selected covariates also modestly improved design sensitivity, but varied by design. Because the mechanisms through which covariance adjustment operate (reducing unexplained variance in the outcome) overlap with those of blocking, the unique contribution of covariance adjustment in terms of improving design sensitivity was tempered by the choice of design. For instance, because blocking on schools in the teacher-randomized design (A) with the geometry outcome eliminates the proportion of outcome variance attributable to districts, adjusting for the covariance between the geometry post- and pre-test provides an additional 9% reduction in variance at the teacher level but supplies no additional reduction at the school level. The additional gains in efficiency due to covariance adjustment are often limited by blocking.

Limitations

We note several limitations in our study. A first limitation is the generalizability of our sample of teachers, schools, and districts. The sample was drawn from teachers, schools, and districts that participated in professional development programs that required an evaluation instrument. Teachers, schools, and districts in the sample likely do not represent the full range of teachers, schools, and districts nationally or the range of those that might participate in professional development. Although this limitation is real, it is important to recognize that samples used to estimate empirical values of design parameters rarely are representative of the samples for a particular study. For instance, even estimates from a nationally representative sample may not be directly applicable to studies using more localized samples for which most studies are conducted (Hedges & Hedberg, 2013). As with all estimates, researchers need to consider carefully to what extent the teacher, school, and district samples upon which the parameters were estimated is applicable to their study sample and be cautious in applying parameter estimates from different subpopulations.

The scope of our analyses was limited to teacher knowledge outcomes. For many studies, there may be other important outcomes. A large portion of professional development programs are designed to enhance teacher knowledge and skills, which in turn is thought to improve classroom teaching and effect gains in student achievement (Yoon et al., 2007). If a teacher fails to acquire new ideas, skills, and knowledge from professional development, it is unlikely that she or he will be able to adapt

classroom instruction and improve student achievement (Yoon et al., 2007). Although it is important that future studies consider other professional development outcome variables, recognize that these variables may also have practical and theoretical limitations. Practically, outcomes such as classroom instruction or value-added scores have been shown to be less reliable (MET, 2012) and thus tend to offer decreased sensitivity for effects. Theoretically, because teacher knowledge often represents a critical mediator, neglecting it may lead to unclear findings (e.g., theory versus implementation failure) and ultimately limit a study's contribution to our understanding of teacher development (e.g., Rossi, Lipsey, & Freeman, 2004).

Similarly, even within the scope of teacher knowledge, the content of a professional development program may be much more targeted than the types of knowledge measured by the mathematical knowledge for teaching measures (e.g., Kennedy, 2016). Alternative knowledge outcomes that are much more targeted to the specific types of knowledge delivered in a specific professional development program may also serve to modify the results. For instance, the observed pre-test–post-test correlations may change in important ways that were not captured here.

Further, our analyses of the estimates considered only a very narrow range of parameter values, assumptions, and plausible designs. For instance, our analyses were limited in terms of cost structures; assumptions about the availability teachers, schools, and districts; and assumptions about effect size variability. As previously noted, our examples were intended to illustrate the use and comparative analysis of designs rather than to comprehensively examine all possible options. Within the context of planning a professional development study, it is important that empirically supported assumptions be made. For instance, in many cases the sample sizes we adopted may be infeasible due to the availability of teachers, schools, and districts. In such cases, one might consider conditional optimal design such that optimal allocations are determined under various fixed sampling constraints (Hedges & Borenstein, 2014).

The limitations presented by the sample, outcomes, and remaining assumptions are particularly relevant for researchers considering using our results to design subsequent studies of professional development. As evidenced by the variability in intraclass correlations and variance explained across outcomes (Jacob, Zhu, & Bloom, 2010), estimates of these parameters can be sensitive to the selected sample and outcomes. Similarly, the results comparing designs are sensitive to these and the additional assumptions. We caution readers to carefully consider potential differences between our analyses and those focused on a given study. Our analyses should be used as an outline of potential considerations rather than a specific blueprint or set of rules for designing professional development studies.

Conclusion

The analyses we have presented provide initial empirical estimates of the parameters needed to design multilevel studies of professional development studies and illustrate the potential of the tradeoffs associated with planning these studies. Clearly, teacher-randomized designs offer the most efficient design. However, for many professional development programs, randomizing at the teacher level will require assigning some teachers to a nontreatment control group. This in turn could serve to undermine the theory of the program by reducing collaboration within schools or districts. Ultimately, the efficiency realized by teacher-randomized designs needs to be weighed against consideration of the threat of study designs that are poorly aligned with the program theory. Collectively, the results suggest a provisional strategy that can be applied in planning professional development studies. A first step might be to identify which units one could sensibly randomize and to evaluate the extent to which randomizing at each level represents a threat to the theory of the professional development program. Once the unit of assignment is outlined and evaluated, a next step is to consider the feasibility of blocking. Considerations include the expectations concerning effect variability and the availability and relative cost of teachers, schools, and districts. As a third step, one should consider the accessibility, cost, and prognostic value of covariates and the extent to which their value overlaps with that of blocking. With the careful and appropriate use of these results, we anticipate that the work will inform the

design of professional development studies and serve as an initial rough guide for delineating empirical and theoretical design considerations for studies of professional development.

Funding

This study was supported by grants from the National Science Foundation (Award numbers 1405601, 1228490, 1552535, and 0927725). The opinions expressed herein are those of the authors and not the funding agencies.

References

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with Mplus. MPlus Web Notes.

Ball, D. L., Thames, M., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407.

Birman, B. F., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., … & O'Day, J. (2007). State and Local Implementation of the No Child Left Behind Act. Volume II-Teacher Quality under NCLB: Interim Report. US Department of Education (ERIC Document Reproduction Service No. ED497970).

Blank, R. K., de las Alas, N., & Smith, C. (2008, January). *Does teacher professional development have effects on teaching and learning? Evaluation findings from programs in 14 states*. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Resources/Publications/Does_Teacher_Professional_Development_Have_Effects_on_Teaching_and_Learning_Analysis_of_Evaluation_Findings_from_Programs_for_Mathematics_and_Science_Teachers_in_14_States.html

Bloom, H. (1995). *Minimum detectable effects in a cluster randomized experiment*. New York, NY: Presentation at the Manpower Development Research Corporation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). 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.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328.

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

Brandon, P. R., Harrison, G. M., & Lawton, B. E. (2013). SAS code for calculating intraclass correlation coefficients and effect size benchmarks for site-randomized education experiments. *American Journal of Evaluation*, *34*(1), 85–90.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. National Commission on Teaching & America's Future, Kutztown Distribution Center, Kutztown, PA.

Desimone, L. M., Garet, M. S., Birman, B. F., Porter, A., & Yoon, K. S. (2003). Improving teachers' in-service professional development in mathematics and science: The role of postsecondary institutions. *Education Policy*, *17*, 613–649.

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

Geyer, J., Davis, M., & Narayan, T. (2016). Intracluster correlation coefficients of household economic and agricultural outcomes in Mozambique. *Evaluation Review*. Retrieved from http://doi.org/10.1177/0193841X16659834

Glassman, J. R., Potter, S. C., Baumler, E. R., & Coyle, K. K. (2015). Estimates of intraclass correlation coefficients from longitudinal group-randomized trials of adolescent HIV/STI/Pregnancy prevention programs. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, *42*(4), 545–53. Retrieved from http://doi.org/10.1177/1090198114568308

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546–582.

Hedges, L., & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics*, *39*, 1–25.

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

Hedges, L. V., & Hedberg, E. C. (2007b). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, *22*(10), 1–15.

Hedges, L., & Hedberg, E. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489.

Hill, H., & Ball, D. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institute. *Journal for Research in Mathematics Education*, *35*, 330–351.

Hill, H., Ball, D., & Schilling, S. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*, 372–400.
Hill, C., Bloom, H., Black, A., & Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.
Hill, H. C., Schilling, S., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*, 11–30.
Hill, H. C., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406.
Institute of Education Sciences. (2012). Research grants request for applications for awards beginning in fiscal year 2013: CFDA Number 84.305A. Washington, DC: U.S. Department of Education, Institute of Education Sciences.
Isaakidis P., & Ioannidis, J. (2003). Evaluation of cluster randomized controlled trials in sub-Saharan Africa. *American Journal of Epidemiology*, *158*, 921–926.
Jaciw, A., Lin, L. & Ma, B. (2014) An Empirical Study of Design Parameters for Assessing Differential Impacts for Students in Group Randomized Trials (June 30, 2014). Retrieved from http://dx.doi.org/10.2139/ssrn.2516005
Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness*, *3*, 157–198.
Jerald, C. (2002). All talk, no action: Putting an end to out-of-field teaching. *The Education Trust*, *9*, 1–14.
Kane, T., Staiger, D. O., & McCaffrey, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Bill and Melinda Gates Foundation. Retrieved from http://eric.ed.gov/?id=ED529895
Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis*, *35*, 370–390.
Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*. Retrieved from http://doi.org/10.3102/0034654315626800
Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review*, *33*(4), 335–357.
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.
Raudenbush, S., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29.
Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. Retrieved from http://doi.org/10.1037/1082-989X.5.2.199
Rossi, A., Lipsey, M., & Freeman, H. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.
Rubin, D. B., & Little, R. J. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: J Wiley & Sons.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. Retrieved from http://doi.org/10.1037/1082-989X.7.2.147
Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.
Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87.
Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*(4), 441–471. Retrieved from http://doi.org/10.3102/1076998610375840
Spybrook, J. (2013). Introduction to special issue on design parameters for cluster randomized trials in education. *Evaluation Review*, *37*(6), 435–444.
Spybrook, J., Ran, S., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education*, *39*(3), 255–267.
Spybrook, J., Westine, C., & Taylor, J. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *1*, 1–15.
Spybrook, J., & Raudenbush, S. W. (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*, 298–318.
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.
Westine, C., Spybrook, J., & Taylor, J. (2014). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*, 490–519.

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.

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(5), 242–278.

Appendix

Example Learning Mathematics for Teaching Assessment Question

Which of the following story problems can be used to represent 1 $^1/_4$ divided by $^1/_2$?	Yes	No
a. You want to split $1^1/_4$ pies evenly between two families. How much should each family get?	1	2
b. You have $1.25 and may soon double your money, How much money would you end up with?	1	2
c. You are making some homemade taffy and the recipe calls for 1 $^1/_4$ cups of butter. How many sticks of butter will you need? (each stick = $^1/_2$ cup)	1	2

Note: The example LMT question is taken from Ball, Thames, and Phelps (2008). The following explanations are given for the correct answers. "The first word problem is division by 2 rather than by $^1/_2$; the second is multiplication by 2 rather than by $^1/_2$ (a subtle yet important point for teaching this content); and the third correctly fits the calculation—using a measurement meaning of division. The important point here, though, is that figuring out which story problems fit with which calculations, and vice versa, is a task engaged in teaching this content, not something done in the solving problems with this content" (p. 400).