Article

Evaluation Review
1-27

DOI: 10.1177/0193841X16665024
erx.sagepub.com

Informing Estimates of Program Effects for Studies of Mathematics Professional Development Using Teacher Content Knowledge Outcomes

Geoffrey Phelps[1], Benjamin Kelcey[2], Nathan Jones[3], and Shuangshuang Liu[1]

Abstract

Mathematics professional development is widely offered, typically with the goal of improving teachers' content knowledge, the quality of teaching, and ultimately students' achievement. Recently, new assessments focused on mathematical knowledge for teaching (MKT) have been developed to assist in the evaluation and improvement of mathematics professional development. This study presents empirical estimates of average program change in MKT and its variation with the goal of supporting the design of experimental trials that are adequately powered to detect a specified program effect. The study drew on a large database representing five different assessments of MKT and collectively 326 professional development programs and 9,365 teachers. Results from cross-classified hierarchical growth models found

[1] Educational Testing Service, Princeton, NJ, USA
[2] University of Cincinnati, Cincinnati, OH, USA
[3] Boston University, Boston, MA, USA

Corresponding Author:
Geoffrey Phelps, Educational Testing Service, 660 Rosedale Rd., Princeton, NJ 08541, USA.
Email: gphelps@ets.org

that standardized average change estimates across the five assessments ranged from a low of 0.16 standard deviations (*SD*s) to a high of 0.26 *SD*s. Power analyses using the estimated pre- and posttest change estimates indicated that hundreds of teachers are needed to detect changes in knowledge at the lower end of the distribution. Even studies powered to detect effects at the higher end of the distribution will require substantial resources to conduct rigorous experimental trials. Empirical benchmarks that describe average program change and its variation provide a useful preliminary resource for interpreting the relative magnitude of effect sizes associated with professional development programs and for designing adequately powered trials.

Keywords
mathematics, professional development, mathematical knowledge for teaching, program evaluation, group randomized trial, program effects

There is a growing consensus that professional development is critical to improving the quality of the teaching workforce (American Federation of Teachers, 2002; Darling-Hammond & Sykes, 1999; National Academy of Education, 2009; The Holmes Group, 1986). Interest in professional development is in part driven by the recognition that teachers need to learn a great deal to achieve the goals set by ambitious new content standards (Ball & Cohen, 1999; Borko, 2004; Putnam & Borko, 1997; Wilson & Berne, 1999). Teacher learning is also seen as a promising way to address large differences that have been observed in teacher effectiveness (see, e.g., Nye, Konstantopoulous, & Hedges, 2004). In most school districts, teachers regularly participate in professional development activities, accounting for a substantial proportion of district expenditure (Miles, Odden, Fermanich, & Archibald, 2004). For example, one recent study of a number of typical school districts found that on average US$18,000 a year was spent per teacher on professional development (TNTP, 2015).

While the investment in teacher professional development is substantial and the need is widely recognized, there is strong evidence that the current system of professional development is not up to the task. Most professional development consists of brief district-led workshops that are disconnected from teachers' work experiences and do not focus directly on the daily challenges of teaching core subjects such as reading and mathematics (Ball & Cohen, 1999; Wilson & Berne, 1999). Summaries of the research base on professional development have found weak evidence and mixed results

(see, e.g., Hill, Beisiegel, & Jacob, 2013; Gersten, Taylor, Keys, Rolfhus, & Newman-Gonchar, 2014).

Given the interest in teachers and their development and a general concern that professional development does little to improve teacher quality, it is no surprise that policy makers, researchers, and educators are calling for reliable empirical evidence that can support the design of effective professional development programs (Desimone, 2009; Garet et al., 2010; Garet et al., 2011; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Major funders have also recognized this need, with the Institute of Education Sciences (IES) having recently funded over 60 projects and an entire program devoted to research on effective strategies for improving teacher quality through professional development (IES, 2012).

To date, evaluations and research on the effects of professional development on teacher learning have been limited by the use of outcomes such as teachers' evaluations of the quality of programs or teachers' self-reports of their knowledge and learning (Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Other outcomes of professional development, such as the quality of instruction or student achievement, present other limitations. Even though student achievement has strong measurement properties, it is arguably too distal because many factors intervene between effective professional development and what students learn (Yoon et al., 2007). Further, since most professional development programs are studied in small-scale evaluations of single site programs, studies that use student achievement outcomes are typically underpowered (Borko, 2004; Wayne et al., 2008).

The absence of suitable direct measures (e.g., appropriate assessments of teacher knowledge) and the challenges in using more distal measures (e.g., student learning) have contributed to a general lack of studies that are adequately designed to detect program effects. For example, in a recent review of 1,300 studies of professional development, only 9 studies had research designs that could support a strong causal inference of the effect of professional development on a relevant outcome measure (Yoon et al., 2007).

In the area of mathematics, a recent and promising alternative to these existing measures are assessments of the specialized types of mathematical knowledge used in teaching the subject (Ball, Thames, & Phelps, 2008; Hill, Schilling, & Ball, 2004; Kersting, 2008; Krauss, Baumert, & Blum, 2008; Phelps, Weren, Croft, & Gitomer, 2014). Measures of mathematical knowledge for teaching (MKT) focus on the content problems that teachers

encounter as they, for example, interpret students' mathematical errors, decide how to represent a concept to best align with an instructional goal, or evaluate mathematics problems to determine the types of challenges student are likely to encounter. Because these assessments focus directly on assessing a type of professional content knowledge, they have received a great deal of attention including, for example, as assessments of teaching quality in studies such as the Measures of Effective Teaching project (Gitomer, Phelps, Weren, Howell, & Croft, 2014), as recommended instruments in studies funded by the National Science Foundation and the Institute of Educational Sciences, and as outcomes used in studies of professional development (Phelps, Kelcey, Jones, & Liu, 2016). This interest is supported by mounting evidence that scores on assessments of MKT are associated with both the quality of mathematics instruction and student learning (Baumert et al., 2010; Hill et al., 2008; Hill, Rowan, & Ball, 2005; Kersting et al., 2010; Kersting et al., 2012; Rockoff, Jacob, Kane, & Staiger, 2011).

Although new assessments of MKT have gained in popularity and are promising for evaluating professional development, there is relatively limited information available to guide researchers in designing rigorous evaluations and experimental trials using these outcomes (Kelcey & Phelps, 2013, 2014). Even basic parameters, such as estimates of the anticipated effect size change in teacher knowledge that results from a professional development intervention, are not readily available. For example, in a recent survey of peer-reviewed studies of professional development, 24 of the 44 studies used teacher knowledge outcomes, but only 9 of these presented statistics that could be used to generate a standardized change in MKT, and none used an experimental design (Phelps et al., 2016).

Gauging Appropriate Effect Sizes to Use in Designing Evaluations of Professional Development

Research from the student literature indicates that developing empirical estimates of program change is an important step in planning future studies, as these estimates often vary substantially by student grade level and other student and school characteristics (Hill, Bloom, Black, & Lipsey, 2008; Lipsey et al., 2012). For example, grade-level differences in students' standardized reading outcomes range from a high of 1.52 standard deviations (*SD*s) for Grades K–1 to a low of 0.06 *SD*s for Grades 11 and 12 (Hill et al., 2008, p. 173). There are also substantial differences by students' background characteristics, school characteristics, and the extent to which tests are broadly or narrowly focused on the content of the treatment. Results

from this line of research indicate that sound study design and interpretation require the use of effect size benchmarks that are tailored to the relevant participant groups and outcomes.

In generating empirical benchmarks, the student outcome literature has relied on summarizing the effect sizes reported across many similar experiments (Lipsey et al., 2012). While this approach is arguably ideal, because it provides a strong basis for accounting for nontreatment effects, such as naturally occurring learning trajectories, it depends on the availability of a large number of prior experiments and is therefore of limited value for new outcomes or emerging areas of interest.

An alternative approach is to describe the distribution of growth or change for a normative sample. For example, the results reported from Hill et al. (2008) above describe differences in students across grade levels without reference to a control group. The implied treatment is the year of instruction and other incidental learning that occurs between test administrations. Arguing for the relevance of such findings, Hill et al. (2008) note that when patterns are "striking and consistent, it is reasonable to use them as benchmarks for interpreting effect size estimates from intervention studies" (p. 174).

Given the limited number of prior experiments using teacher knowledge as an outcome, a similar approach for teacher professional development (where one would examine normative patterns in change across many professional development programs) is warranted. In the case of teacher learning, estimates of change are assumed to result from the program treatment between test administrations. The implied control comparison is the group of teachers who did not receive professional development. This comparison relies on the assumption that teachers who did not receive professional development would not show change on the knowledge assessment. There is some emerging evidence that supports the basic assumption that teacher knowledge—absent a professional development treatment—remains relatively stable over time.[1] Although parameters generated from descriptions of longitudinal data cannot account for any growth that occurred due to factors other than the professional development treatment, they are arguably much better than alternatives such as using parameters from the student literature or from a research context in an entirely different field (e.g., Jacob, Zhu, & Bloom, 2010).

In this study, we drew on a large data set that collectively represents 326 professional development programs with 9,365 participating teachers for five different assessments of MKT. Having such a large number of programs allows us to generate estimates of average program-level change in

teacher knowledge and its variation across programs. Because professional development often focuses on groups of teachers with particular background characteristics (e.g., beginning teachers, teachers at a particular grade level, teachers in high-poverty schools), the data also provided opportunities to probe differences in change for teacher groups.

The remainder of the article is organized as follows. The method section describes the data set and how we estimated average program change in MKT. Results are presented in two sections. In the main analysis, we used growth models to generate change estimates for the professional development programs in the sample for each of the five assessments and for relevant teacher groups. Next, we used the change estimates to illustrate necessary sample sizes of schools and teachers to adequately power school-randomized trials. We conclude by discussing the main findings, limitations that need to be considered, and implications for designing rigorous studies of professional development.

Method

The data used in this analysis were collected through the Teacher Knowledge Assessment System (TKAS). TKAS is an online platform used to administer assessments of MKT developed through the Learning Mathematics for Teaching project at the University of Michigan (http://sitemaker.umich.edu/lmt/home). The TKAS assessments are made available to trained users who choose to administer pairs of equated forms (i.e., test forms with different sets of questions that have been statistically adjusted to have the same difficulty) either as a fixed set of questions or as an adaptively selected set of questions based on participant performance (i.e., computer adaptive testing).

Instruments

Our analyses focused on five assessments: Elementary Number Concepts and Operations; Elementary Patterns, Functions, and Algebra (EL PFA); Middle School Number Concepts and Operations (MS NCOP); Middle School Patterns, Functions, and Algebra (MS PFA); and Grades 4–8 Geometry (4–8 GEO). Each of these assessments included multiple parallel forms designed to focus on the same content. These parallel forms were calibrated in fall 2010 taking advantage of the balanced random assignment of the forms at the pretest administration. Random assignment item response theory (IRT) form equating (Kolen & Brennan, 2004) was used to adjust the score outcomes to the same difficulty metric. Additional

technical information is available on the project website (http://sitemaker.umich.edu/lmt/home).

Sample

The sample used for this study included participants who completed assessments after July 2010 (the date when TKAS began to administer background surveys) through March 2013. The final sample for each assessment is shown in Table 1. The last column presents the same descriptive statistics for a nationally representative sample of teachers drawn from the 2011 to 2012 Public Elementary/Secondary School Universe Survey Data, which is part of the Common Core of Data from the National Center for Education Statistics.

Table 2 summarizes the number of programs, schools, and teachers for each assessment. Even though the number of schools is on the order of 10 times the number of programs, schools were not uniquely nested within programs. There was considerable cross classification with multiple programs nested within a given school. Table 2 also indicates substantial missing data on the posttest assessment from a high of 60% for EL PFA to a low of 46% for MS NCOP. For each of the growth models described below, we estimated parameters using restricted maximum likelihood to accommodate missing data under the missing at random assumption (Allison, 2001).

Analytic Methods

Models. To estimate the average change in teacher knowledge across programs, we fit a series of models designed to capture the average pre- to postgains in MKT. In our models, we approximated the empirical distribution of program effect sizes or the anticipated difference between control and experimental units. To approximate the distribution of program effect sizes, we employed growth curve models to replicate the conditions under which meta-analyses would be conducted using experimental data. We do this by summarizing multiple estimates of program change to approximate an analysis of effect size estimates.

An important assumption supports the use of change score estimates as the basis for the effect sizes required for experimental design. We assume that teachers' content knowledge is relatively stable in the absence of exposure to substantial professional development. Unlike students who receive formal learning opportunities across an entire school year, teachers' formal

Table 1. Descriptive Statistics of Teachers for Each Outcome Measure.

Teacher Characteristics	Outcome Measures					
	EL NCOP (n = 3,340, %)	EL PFA (n = 2,074, %)	MS NCOP (n = 1,020, %)	MS PFA (n = 2,203, %)	4–8 GEO (n = 728, %)	CCD (n = 37,497, %)
White	87	85	76	68	81	90
Female	92	93	77	77	88	76
NBPTS mathematics	2	2	13	18	8	1
Mathematics degree	7	7	37	41	23	5
Grade taught[a]						
K–2	44	43	5	4	27	29
3–5	58	56	14	11	40	32
6-8	8	10	60	66	27	29
9–12	1	1	32	33	21	33
Years taught						
0–3	20	20	25	23	16	11
4–15	53	54	54	57	58	53
>15	27	26	21	20	26	36
Percentage of free and reduced lunch						
0–0.25	14	16	9	10	14	24
0.26–0.50	32	28	32	24	34	28
0.51–0.75	36	38	39	41	39	26
0.76–1	18	18	20	25	13	22

(continued)

Table 1. (continued)

Teacher Characteristics	Outcome Measures					
	EL NCOP (n = 3,340, %)	EL PFA (n = 2,074, %)	MS NCOP (n = 1,020, %)	MS PFA (n = 2,203, %)	4–8 GEO (n = 728, %)	CCD (n = 37,497, %)
Region						
Northeast	1	0	5	2	0	21
South	49	74	60	80	64	39
Midwest	23	14	10	4	17	22
West United States	27	13	25	14	19	18
Urban						
City	21	17	31	37	24	28
Suburb	25	27	17	14	23	32
Town	14	15	15	17	15	12
Rural	40	41	37	32	38	27
Program type						
Summer	17	16	3	6	0	—
School year	32	32	38	39	30	—
Whole year	51	52	59	55	70	—

Note. Summary statistics from Common Core of Data (CCD) are presented to aid in interpreting how representative the data are for each TKAS outcome. The CCD data are presented for teachers in Grades K–12. EL NCOP = Elementary Number Concepts and Operations; EL PFA = Elementary Patterns, Functions, and Algebra; MS NCOP = Middle School Number Concepts; MS PFA = Middle School Patterns, Functions, and Algebra; NBPTS = National Board for Professional Teaching Standards; 4–8 GEO = Grades 4–8 Geometry.

[a]Percentage of grades taught do not add up to 100 as multiple options could be selected.

Table 2. Program, School, and Teacher Sample Sizes for Each Outcome Measure.

	Outcome Measures				
N.	EL NCOP	EL PFA	MS NCOP	MS PFA	4–8 GEO
Program	104	53	47	82	40
School	995	613	503	1,093	349
Teacher	3,340	2,074	1,020	2,203	728
Pretest	3,336	2,070	1,019	2,202	726
Posttest	1,436	822	550	1,174	392

Note. EL NCOP = Elementary Number Concepts and Operations; EL PFA = Elementary Patterns, Functions, and Algebra; MS NCOP = Middle School Number Concepts; MS PFA = Middle School Patterns, Functions, and Algebra; 4–8 GEO = Grades 4–8 Geometry.

learning opportunities are typically extremely short in duration, insubstantial, or simply not present (Wei, Darling-Hammond, Andree, Richardson, & Orphanos, 2009), suggesting that teachers are not developing MKT on their own or in the limited learning opportunities they typically receive. Absent a formal professional development intervention, little or no change in knowledge is likely to occur. Another argument for using change score estimates is practical. Since there is no large pool of well-executed experimental professional development studies to outline the distribution of effect sizes, it makes sense to rely on existing data to conduct an initial investigation through change score models.

To implement these analyses, we drew on the following average gain score[2] models using multivariate cross-classified random effects models:

$$\begin{aligned} Y_{ijk1} &= \mu_1 + r_{j1} + u_{k1} + \varepsilon_{ijk1}(\varepsilon_{ijk1}, \varepsilon_{ijk2}) \sim MVN(0, \Sigma_\varepsilon), (r_{j1}, r_{j2}) \sim MVN(0, \Sigma_r), \\ Y_{ijk2} &= \mu_2 + r_{j2} + u_{k2} + \varepsilon_{ijk2}(u_{k1}, u_{k2}) \sim MVN(0, \Sigma_u). \end{aligned} \quad (1)$$

We use Y_{ijk1} and Y_{ijk2} as the pre- and posttest MKT scores for teacher i in program j in school k, μ_1 and μ_2 as the conditional average pre- and posttest scores, r_{j1} and r_{j2} as pre- and posttest program-specific random effects, u_{k1} and u_{k2} as pre- and posttest school-specific random effects, and ε_{ijk1} and ε_{ijk2} as the pre- and posttest teacher-specific residuals. Further, each pair of random effects was set to follow a multivariate normal distribution with estimated variances ($\sigma^2_{.1}, \sigma^2_{.2}$) and covariance ($\sigma^2_{.12}$). In turn, we estimated parameters using restricted maximum likelihood under the assumption that data were missing at random.

We used $\mu_2 - \mu_1$ as the average change score from pre- to posttest across all professional development programs. To establish standardized measures of change for school-randomized designs, we standardized the gains on the basis of full unconditional teacher and school posttest variance. Similarly, to describe the variance of change scores across programs, we used the variance of the differences or $\sigma^2_{r1} + \sigma^2_{r2} - 2\sigma^2_{r12}$.

To describe the extent to which changes in teacher knowledge varied by teacher, school, and program characteristics, we introduced covariates one at a time. The six covariates we considered included (1) mathematics degree, (2) teaching experience (0–3, 4–15, and >15 years), (3) school percentage of students eligible for free/reduced lunch (0–0.25, 0.26–0.50, 0.51–0.75, and 0.76–1), (4) geographical region of school (Northeast, South, Midwest, and West United States), (5) urbanicity of school (city, suburb, town, and rural),[3] and (6) professional development program length type (summer, school year, and year-round).

Implications for School-Randomized Designs

To illustrate the practical implications associated with our empirical results, we estimated the sample sizes necessary to sufficiently power school-randomized studies under several conditions. To estimate these sample sizes, we drew on the effect size estimates developed in the current analyses as well as variance components previously reported in the literature for these measures under a pretest adjusted design (Kelcey & Phelps, 2014). We specified the two-level school-randomized experiments with teachers nested within schools as:

$$\begin{aligned} Y_{ij}^{(A)} &= \pi_{0j} + \pi_1(X_{ij} - \bar{X}_j) + \varepsilon_{ij} \quad \varepsilon \sim N(0, \sigma_1^2) \\ \pi_{ij} &= \beta_{00} + \delta T_j + \beta_1 \bar{X}_j + u_{0j} \quad u \sim N(0, \sigma_2^2). \end{aligned} \tag{2}$$

Here, $Y_{ij}^{(A)}$ as the outcome for teacher i in school j for outcome A, π_{0j} as the school-specific intercept, X_{ij} as the pretest with $\bar{X}_{ij}$ as its school mean, ε_{ij} as the teacher-specific error, β_{00} as the overall mean, T_j as the treatment indicator, π_1 and β_1 as the coefficients, and u_{0j} as the school-specific random effect.

To estimate the sample sizes necessary to sufficiently power a study, we considered fully balanced designs with 80% power under a two-tailed test with a Type I error rate of 0.05. Using the standardized average program change estimates and the estimates taken from Kelcey and Phelps (2014), we estimated sample sizes for school-randomized trials that assumed 2, 4, 6, or 12 teachers within a school.

Table 3. Average Pre- to Posttest Change for Each Outcome Measure.

Outcome Measures	Standardized average change	Variance of average change across programs	Program	School	Teacher
Elementary Number Concepts	.23	.05	104	995	3,340
Elementary Algebra	.18	.07	53	613	2,074
Middle School Number Concepts	.21	.10	47	503	1,020
Middle School Algebra	.16	.06	82	1,093	2,203
Grades 4–8 Geometry	.26	.14	40	349	728

Note. The standardized average change is calculated using the estimated average change score divided by the square root of school-level and teacher-level variance for posttest scores. Raw estimates and associated standard errors are available upon request.

Results

Average Change in Teacher Knowledge

We begin by describing standardized average change in teacher knowledge across all professional development programs as well as the variability of these change scores across programs (Table 3). The standardized average change estimates across the five assessments ranged from a low of 0.16 *SD*s for MS PFA to a high of 0.26 *SD*s for Grades 4–8 GEO. Relative to the magnitude of the average change, the variance across programs tended to be sizable (Table 3). Take, for example, the MS NCOP assessment. Its standardized average change was 0.21, but the variance was 0.10. Assuming the average changes are roughly normally distributed across programs, this would suggest that although the average change was 0.21, the average change for programs at the 95th percentile, for example, might reach as high as $0.21 + 2\ (0.31) = 0.83$.

Change by Teacher Groups

We next examined average change by different teacher subgroups. We focused on teacher groups that can often be readily identified through information commonly available to researchers either during the design or analysis phase of the study. These include characteristics of teachers, characteristics of a teacher's school, and basic characteristics of the

professional development program.[4] Entering these covariates independently into each analysis provides a useful set of benchmarks for gauging whether these teacher group differences influence program change.

The results are presented in Table 4.[5] While some of the observed differences between groups are quite sizable, we are cautious to note some comparisons derive from relatively small sample sizes and these differences may be susceptible to random error. For this reason, our main interest was in identifying patterns that are replicated across the five assessments.

Only a small number of tests of difference for the teacher groups showed a significant change in teacher knowledge. Furthermore, for most teacher groups, there was not a consistent pattern observed for more than two of the assessments. The one exception was mathematics degree, where across all five assessments, teachers with a mathematics degree had a larger change in knowledge compared to teachers without a mathematics degree. However, for mathematics degree, the difference was only significant for one of the five of the assessment outcomes. Given the lack of interpretable difference observed for teacher groups, we focus in the remainder of the results on the overall average change for each outcome.

Implications for Study Design

In this section, we illustrate how the estimates of average change can aid in study design and in interpretation of study results. We begin by describing variation in change estimates across programs. We select different points in the distribution of program change to suggest empirical benchmarks of different magnitudes. Next, we draw on these empirical benchmarks to estimate the sample sizes needed to detect a significant program change for school-randomized trials at the mean of the standardized distribution of program change and at $+0.5$ and $+1$ *SD*s above the mean.

Describing variation in program change. In Figure 1, we plotted the variation in standardized program change for each assessment.[6] There was substantial difference in the deviation of program change, with programs in MS NCOP deviating far less than programs in 4–8 GEO.

To make these distributions of change more useful for the purpose of study design and interpretation, we described the magnitude of change at -1.0, -0.5, 0, $+0.5$, and $+1.0$ *SD*s in the distribution of program change (Table 5).[7]

There are a number of general trends of note that are illustrated in Table 5. For all of the assessments, there was a sizable proportion of programs that

Table 4. Standardized Average Change by Teacher Groups for Each Outcome Measure.

	Outcome Measures				
Teacher Groups	EL NCOP	EL PFA	MS NCOP	MS PFA	4–8 GEO
Overall change	.23	.18	.21	.16	.26
Mathematics degree					
No	.23	.17	.17	.15	.25
Yes	.28	.29*	.27	.18	.26
Teaching experience					
0–3	.25	.20	.14	.07	.43
15–Apr	.25	.20	.23	.21*	.33
>15	.17	.13	.22	.10	−.03**
Percentage of free reduced lunch					
0–0.25	.16	.00	*.45*	.16	.57
0.26–0.50	.24	.16	.24	.10	.28
0.51–0.75	.25	.21	.19	.16	.11*
0.76–1	.21	.24	.06	.22	.39
Region					
Northeast	.39	—	—	.27	—
South	.27	.17	.38	.22	.27
Midwest	.30	.19	*.16*	−.07	.34
West United States	.13	.18	−.04	.00	.18
Urban					
City	.26	.18	.32	.13	.37
Suburb	.22	.08	.06	.06	.38
Town	.19	.24	.21	.19	.01*
Rural	.22	.20	.21	.19	.25
Program type					
Summer	.25	−.14	*.18*	.23	—
School year	.22	.26	.24	.16	.42
Whole year	.21	.20	.20	.15	.18

Note. Change estimates for teacher groups in italics are based on samples of less than 50 teachers. The standardized average change is calculated using the estimated average change score divided by the square root of school-level and teacher-level variance for posttest scores. Raw estimates and associated standard errors are available upon request. The first category of each teacher group variable is used as the reference group for testing differential average change across subgroups. EL NCOP = Elementary Number Concepts and Operations; EL PFA = Elementary Patterns, Functions, and Algebra; MS NCOP = Middle School Number Concepts; MS PFA = Middle School Patterns, Functions, and Algebra; 4–8 GEO = Grades 4–8 Geometry.
*$p < .05$. **$p < .01$. ***$p < .001$.

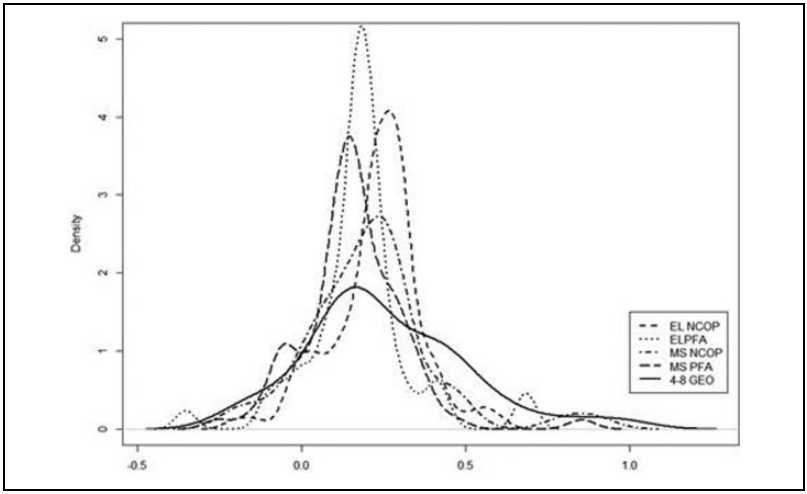

Figure 1. Distribution of program change for each assessment. Density plots are produced based on the (shrunken) empirical Bayes estimates of the average change for each program. Elementary Number Concepts, Elementary Algebra, Middle School Number Concepts, Middle Algebra, and Grades 4–8 Geometry.

Table 5. Change in Knowledge at Different Quantiles in the Standardized Distribution of Program Change for Each Outcome Measure.

	Standardized Distribution				
Outcome Measures	−1 *SD*	−0.5 *SD*	Mean	+0.5 *SD*	+1 *SD*
Elementary Number Concepts	.00	.12	.23	.34	.46
Elementary Algebra	−.09	.04	.18	.32	.45
Middle School Number Concepts	−.10	.06	.21	.36	.52
Middle School Algebra	−.09	.04	.16	.28	.41
Grades 4–8 Geometry	−.12	.07	.26	.45	.64

Note. The standard deviation of program-level change for each assessment is the square root of the program-level variance for change score divided by the square root of school-level and teacher-level variance for posttest score.

had a very small or no positive change in teacher knowledge. This is particularly apparent for the PFA assessments where approximately 30% of the programs showed no change. It is also noteworthy that programs at an *SD* above the mean in the program distribution showed a substantial

Table 6. Unconditional Interclass Correlation Coefficient (ICC) and Confidence Intervals and Proportion of Variance Explained by Pretest for Each Outcome Measure.

	ICC			Variance Explained	
Outcome Measures	Estimate	Low	High	Teacher	School
Elementary Number Concepts	.22	.16	.29	.32	.65
Elementary Algebra	.31	.23	.40	.24	.65
Middle School Number Concepts	.35	.21	.47	.34	.76
Middle School Algebra	.16	.07	.27	.35	.59
Grades 4–8 Geometry	.26	.19	.33	.27	.83

Note. Low is the lower bound and high is the upper bound of the 95% bootstrapped confidence interval for the ICC. Estimates are taken from analyses using the TKAS data set presented in Kelcey & Phelps (2014, table 2, p. 17 and table 6, p. 19).

average change in teacher content knowledge ranging between 0.41 and 0.64 *SD*s across the five assessments.

Power estimates for school-randomized trials. The empirical estimates for change are a helpful guide in making decisions about the sample size needed to adequately power a study. In this section, we consider the sample size implications for school-randomized trials. We used the program estimates generated above in Table 5 for the mean, +0.5 *SD*, and +1 *SD*. These correspond to the points on the distribution of program change, where 50% of programs perform better, 30.9% perform better, and 15.9% perform better. Our estimates of the intraclass correlation coefficient and variance components for each outcome were drawn from the TKAS study under a previous wave of data (Kelcey & Phelps, 2014; Table 6).[8] We then present the sample that would be required for a school-randomized trial for each of the five outcomes at each of these points in the program distribution (Table 7). We estimated the teacher and school sample based on a design, where there are 2, 4, 6, or 12 teachers within each school.

Regardless of the test outcome, programs that are designed to achieve a significant change at the mean of program distribution may require large samples of teachers. The sample requirements, when we assume there are 4 teachers in each school, range from a low of 364 teachers in 91 schools for 4–8 GEO to a high of 936 teachers in 234 schools for EL PFA. Few professional development programs are conducted at this scale. However, the sample size estimates for detecting a significant change at an *SD* above

Table 7. Number of Schools Needed to Detect a Significant Change in Teacher Knowledge for School-Randomized Trials Estimated at Different Quantiles of the Program Distribution.

	Mean				+0.5 *SD*				+1 *SD*			
	Teachers Within School				Teachers Within School				Teachers Within School			
Outcome Measures	2	4	6	12	2	4	6	12	2	4	6	12
EL NCOP	205	126	100	74	95	59	47	35	53	33	27	20
EL PFA	361	234	192	150	116	76	62	49	60	76	62	49
MS NCOP	243	146	113	81	84	51	40	29	41	26	20	15
MS PFA	387	221	167	111	128	74	56	38	61	36	27	19
Grades 4–8 Geometry	141	91	74	58	48	32	26	21	25	17	14	12

Note. Total teacher sample can be calculated by multiplying the teachers within school by the number of schools. EL NCOP = Elementary Number Concepts and Operations; EL PFA = Elementary Patterns, Functions, and Algebra; MS NCOP = Middle School Number Concepts; MS PFA = Middle School Patterns, Functions, and Algebra; 4–8 GEO = Grades 4–8 Geometry.

the mean of the program distribution are more reasonable, ranging from 68 teachers in 17 schools to a high of 304 in 76. Recognizing that only half the sample will participate in a treatment condition, this sample size is well within the capabilities of many professional development programs. Researchers choosing to power a design at an *SD* above the mean of the program distribution, however, would need to have great confidence in the strength of the program treatment to decide that their program will outperform 84.1% of the programs represented in the TKAS sample.

Discussion

The use of assessments of MKT as outcomes for evaluating mathematics professional development has developed rapidly over the past decade. Arguably, among the best evidence for this interest comes from the data set used in this study. In just a three-year period, over 325 programs used the MKT assessments in pre- and posttest administrations. While, only a few studies have been published in the peer-reviewed literature since these assessments became available, the high level of interest suggests that teacher knowledge outcomes have a central role to play in the evaluation and improvement of mathematics professional development. For these assessments to be useful, it will be critical that studies are designed that

are adequately powered to detect program effects. In this section, we highlight key findings from this analysis, discuss limitations that need to be taken into account when applying the empirical benchmarks, and discuss how research on the use of teacher knowledge outcomes can contribute to calls for a more comprehensive and rigorous program of professional development research.

For each of the five assessments, we observed substantial variation across programs in average change in teacher knowledge. There are many possible explanations that might explain this variation including program length and intensity, content focus and rigor, and the quality of the instruction provided to teachers. Because the focus of our analysis was on providing estimates that could guide researchers in the design of adequately powered studies, we focused on characteristics of teachers and programs that are often available to researchers at the design phase of a study. The only characteristic that was associated with higher levels of average change in teacher knowledge across all five assessments was having a mathematics degree. While the analysis we conducted does provide some guidance on study design, it does not resolve the larger question of what might explain the substantial variation in program change. This is an area of inquiry that will be important to pursue in future studies.

The analyses also revealed substantial variation in program change across the five assessments. Depending on the outcome assessment used, the standardized average program change ranged from 0.16 for MS PFA to 0.26 for 4–8 GEO. To put this into perspective, adequately powered studies for a school-randomized trial would require 221 and 91 schools and a total of 884 and 364 teachers, respectively, for these two assessments (assuming that there are 4 teachers in each school). This strongly suggests that researchers interested in designing adequately powered studies need to attend to the specific knowledge outcome that will be used.

Finally, these findings can also be considered in respect to the analogous lines of research using student outcomes. There are both interesting differences and similarities across these lines of work. One notable difference is how change is associated with relevant groups. In the student literature change estimates vary by student demographic characteristics, school characteristics, and other features of students such as grade level (Hill et al., 2008; Lipsey et al., 2012). In contrast, there is limited evidence from our analysis that differences in program-level change are associated with teacher groups. But there are also notable similarities for the two lines of work. Research on students has found that change estimates vary by the outcome according to subject matter differences and the degree to which the

outcome has a broad or narrow topic focus. We see similar variation in teacher knowledge outcomes associated with both the mathematics topic and grade level of the assessment.

Arguably, the most striking and important similarity between the two lines of work is the general finding that there is no single "one-size-fits-all" change estimate that can be applied universally across participant groups and assessment outcomes. Studies focused on student outcomes have demonstrated that general rule of thumb effect size benchmarks, such as Cohen's *d*, are potentially misleading (Schochet, 2008). These general benchmarks do not reflect the variation observed across groups and outcomes and often appear to provide estimates for effect sizes that are larger than those observed in educational experiments. With respect to teacher knowledge outcomes, general benchmarks may also set unreasonable high standards for change. Recall that the suggested effect size magnitudes for Cohen's *d* are small = 0.2, medium = 0.5, and large = 0.8 (Cohen, 1988). Across the teacher content knowledge assessments, the general trend is toward substantially smaller benchmarks than those indicated by Cohen's *d*. Results from both the student literature and an emerging set of findings focused on teacher knowledge outcomes suggest that empirical benchmarks provide more realistic targets that in turn allow for more defensible design decisions and empirically grounded interpretations of research results.

Limitations

The data in TKAS are largely populated by evaluations of professional development programs, where the goal is to detect a pre- to posttest change in teacher content knowledge. Because these studies do not typically include a control group, there is no way to determine whether observed changes are due to participation in the professional development program or to other factors not associated with the treatment. While basic pre- to posttest change parameters are useful for the purpose of describing the variation in change across programs, they do not allow for the stronger claim that the observed change is the effect of a program treatment (Lipsey et al., 2012). Threats, such as not accounting for natural growth trajectories, could lead to a general bias toward overestimates of true program effects. It is important to emphasize, however, that current research on teacher learning has consistently found that teachers demonstrate low levels of MKT with limited evidence that this knowledge changes across a teacher's career. This suggests that substantial natural growth is unlikely absent strong interventions.

Another concern is that the TKAS data do not provide detailed descriptions of the professional development programs. Our analysis assumes that all of the programs in TKAS are treatments designed to improve teacher content knowledge. The observed variation represents differences in program intensity, quality of implementation, and so forth. However, it is possible that some proportion of the programs that have been included are not "programs" at all, but rather groups of teachers who are assigned to participate in studies as nontreatment controls.

We also observed a large number of programs for all assessments that showed a negative average change. There are few possible explanations. It is possible, for example, that some programs may lead to an increase in confusion over the mathematics tested by the assessment outcomes. Especially when ideas are new or complex, this can lead to an initial drop in performance as learners work to consolidate their understanding. While this could explain a small drop in performance, we think it is more likely that negative average change is due to systematic difference between the pretest and posttest condition. One possible explanation is that teachers are more motivated at pretest to carefully complete all questions than at posttest. This change in motivation could be due to many factors including test takers hurrying to complete a test at the end of the last program session, reduced interest in the test at a second viewing, or even reduced effort that might result from a negative experience in a challenging professional development program. All of these factors could depress gain scores.

There is substantial missing data for all assessments. The growth models used in the analysis allowed for estimating missing values on postassessment using the maximum likelihood function. This guards, to some extent, against the concern that teachers missing at posttest could systematically bias the results. As a further check, all models were run using just the restricted sample of teachers with complete pre- and posttest data. Whereas this analysis did reveal that teachers missing posttests had significantly lower scores at pretest than teachers with complete data, there was no significant difference in the program change estimates from those reported in this article (Phelps et al., 2016). This suggests that the missing data are not systematically biasing the change estimates.

Another limitation stems from the programs that are represented in TKAS. These are likely among the most ambitious of professional development programs with a commitment to evaluation and improvement. Even though the sample is large, and the assessment outcomes are among the most used teacher knowledge assessments represented in the literature,

these programs and the associated outcomes may not be representative of professional development in general.

Arguably, the most reasonable approach to generating empirical benchmarks, especially in situations where measures and technologies are relatively new, is to work with existing samples of studies to generate a provisional set of empirical benchmarks. These benchmarks should then be regularly updated as more comprehensive and ideally higher quality data become available. This process of updating would eventually include other valuable information about the substantive importance of the magnitude of change estimates of teacher content knowledge. For example, the rapidly expanding use of teacher knowledge outcomes to evaluate professional development will likely include studies designed to investigate how changes in teacher knowledge mediate the effects of professional development on both instruction and student outcomes. Findings from such studies would provide additional important guidance useful for selecting effect size estimates that are substantively large enough to lead to meaningful differences in student learning. We see the current study results as an important preliminary step in this direction.

Conclusion

Professional development is increasingly seen as a critical lever for improving teacher quality and ultimately student achievement (Desimone, 2009). However, efforts to study and evaluate professional development have been curtailed by a lack of suitable outcomes that measure the knowledge and skills that are typically the focus of professional development. Researchers have defaulted to a variety of proxies, including reliance on teachers' self-reports of their knowledge and learning. This approach relies on suspect measures that may have little relationship to what teachers actually learn and are able to do in their teaching practice. Or researchers have decided to ignore teacher learning altogether, instead focusing on more distal outcomes such as student learning. This approach tends to treat professional development as a black box, ignoring the mechanisms that underlie how professional development works to influence teaching and student learning across varied contexts.

A new generation of measure of MKT has recently been developed and has quickly been adopted for evaluation of professional development programs. The results presented in this analysis, however, suggest that careful consideration needs to be given to the appropriate use of MKT assessments. School-randomized trials, for example, will require samples in the hundreds

of teachers to be adequately powered to detect the average change realized by professional development programs using these assessments. Indeed, these studies will need to be so large that only a few districts in the United States would have a sufficient number of schools to achieve an adequately powered design. And even for studies designed to detect an effect that is one *SD* above the program average, the samples and study size will require substantial fiscal and human resources. Such studies will likely need to be reserved for a limited number of programs that have already demonstrated promise through a series of more modest exploratory studies. This general finding is of direct relevance to funding agencies making decisions on the best ways to support research on professional development.

Borko (2004) has argued for a unified program of research on professional development that includes a combination of smaller studies designed to investigate a single program often at a single site and larger scale studies investigating the implementation of programs with multiple providers across multiple contexts. Implicit in these arguments is a need for research designs and associated measures that are appropriate for both small-scale and large-scale designs. Assessments of teacher content knowledge can play an important role in this vision. Although the empirical benchmarks we have generated suggest that small-scale experimental trials focused on single programs at a single site will be underpowered, larger scale trials of mathematics professional development are potentially feasible using teacher knowledge outcomes. Developing a better understanding of both the potential and limitations of teacher knowledge outcomes is an important step toward clarifying the role that these measures can play in improving the rigor and quality of professional development research, teachers' learning opportunities, and, ultimately, the quality of mathematics instruction.

Authors' Note

The opinions expressed herein are those of the authors and not the funding agency.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by grants from the National Science Foundation (Award Nos. 0927725 and 1405601).

Notes

1. The assumption of limited natural growth in teacher content knowledge is generally supported by reviews of professional development that show limited or mixed results (see, e.g., Hill et al., 2013; Gersten et al., 2014, Yoon et al., 2007). A stronger test, however, would be to examine knowledge growth for teachers not participating in professional development. There is emerging evidence from at least one large-scale–randomized trial funded by the National Science Foundation (investigating the effect of professional development, mathematical knowledge for teaching, and instruction on student outcomes, Award #0918383) that teachers assigned to the nontreatment control group showed no significant change in knowledge over the course of the study (R. Jacob, personal communication, June 3, 2016).
2. An alternative approach would be to use covariate-adjusted mean scores. However, it is not clear how such estimates would relate to the types of treatment effects and experimental designs we are aiming to inform.
3 We followed the region categories defined in the National Assessment of Educational Progress. For details of the definition of region, visit http://nces.ed.gov/nationsreportcard/glossary.aspx?nav=y#region
4. In exploratory analyses, we did not find differences across the five outcomes by program duration or intensity and have therefore chosen to not include these variables in this article. For additional information on these analyses, please refer to Phelps, Kelcey, Jones, and Liu (2016).
5. Full tables for the estimates of change for teacher groups are available from the corresponding author upon request.
6. Density plots were produced based on the best linear unbiased predictor of the pre to posttest program-level change estimated using the cross-classified models described in the method section. The plot shown is for the first imputed database. There was little difference in the observable characteristics of the curves across the five imputed databases.
7. As indicated in Figure 1, all distributions are roughly normal which generally supports the transformation and interpretation of a standardized distribution of change.
8. Estimates are taken from Kelcey & Phelps, 2014, Table 2, p. 17 and Table 6, p. 19.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

American Federation of Teachers. (2002). *Principles for professional development: AFT's guidelines for creating professional development programs that make a difference*. Washington, DC: Author.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice based theory of professional education. In L. D. Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco, CA: Jossey-Bass.

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

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180. doi:10.3102/0002831209345157

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Darling-Hammond, L., & Sykes, G. (Eds.). (1999). *Teaching as the learning profession: Handbook of policy and practice*. San Francisco, CA: Jossey-Bass.

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

Desimone, L. M., Porter, A. C., Garet, M., Yoon, K. S., & Birman, B. (2002). Does professional development change teachers' instruction? Results from a three-year study. *Educational Evaluation and Policy Analysis*, *24*, 81–112.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., & Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (U.S. Department of Education Report NCEE 2011-4024). Washington, DC: U.S. Department of Education.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., & Doolittle, F. (2010). *Middle school mathematics professional development impact study: Findings after the first year of Implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches* (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education

Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

Gitomer, D., Phelps, G., Weren, B., Howell, H., & Croft, A. (2014). Evidence on the validity of content knowledge for teaching assessments. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 493–528). San Francisco, CA: Jossey-Bass.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*, 476–487.

Hill, H. 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., Blunk, M. L., Charalambous, Y., Lewis, J. M., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

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. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406. doi:10.3102/00028312042002371

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

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.

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.

Kelcey, B., & Phelps, G. (2014). Strategies for improving power in school randomized studies of professional development. *Evaluation Review*, *37*, 520–554.

Kersting, N. B. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*, 845–861.

Kersting, N. B., Givvin, K., Sotelo, F., & Stigler, J. W. (2010). Teacher's analysis of classroom video predicts student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, *61*, 172–181.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*, 568–589. doi:10.3102/0002831212437853

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *ZDM*, *40*, 873–892.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W. , . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: U.S. Government Printing Office. Retrieved from http://ies.ed.gov/ncser/pubs/20133000/

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance*, *30*, 1–26.

National Academy of Education. (2009). *Teacher quality* (S. M. Wilson, Ed.). Washington, DC: Author.

Nye, B., Konstantopoulous, S., & Hedges, L. V. (2004). How large are teacher effects? *Education Evaluation and Policy Analysis*, *26*, 237–257.

Phelps, G., Weren, B., Croft, A., & Gitomer, D. (2014). *Developing content knowledge for teaching assessments for the Measures of Effective Teaching study* (ETS Research Report No. RR-14 33). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12031

Phelps, G., Kelcey, B., Jones, N., & Liu, S. (2016). *Implications for using teacher content knowledge outcomes to evaluate mathematics professional development*. Princeton, NJ: Educational Testing Service Research Report.

Putnam, R., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *The international handbook of teachers and teaching* (pp. 1223–1296). Dordrecht, the Netherlands: Kluwer.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*, 43–74. doi:10.1162/EDFP_a_00022

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

TNTP. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. Washington, DC: Authors.

The Holmes Group. (1986). *Tomorrow's teachers: A report of the Holmes Group*. East Lansing, MI: The Holmes Group.

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*, 469–479.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX: National Staff Development Council.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education*, *24*, 173–209.

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.

Author Biographies

Geoffrey Phelps is a senior research scientist at the Educational Testing Service. His research focuses on the measurement of teacher content knowledge, teacher development, and the relation of teacher knowledge and teaching quality.

Benjamin Kelcey is an associate professor of Quantitative Research Methodologies at the College of Education, Criminal Justice, and Human Services at the University of Cincinnati. Kelcey's research focuses on causal inference and measurement methods within the context of multilevel and multidimensional settings such as classrooms and schools.

Nathan Jones is an associate professor of Special Education at Boston University. Jones' research focuses on teacher quality, teacher development, and school improvement, with a specific emphasis on the use of measures of teacher effectiveness in evaluation systems.

Shuangshuang Liu is a research associate at the Educational Testing Service and is also a PhD candidate at Teachers College, Columbia University.