Symposium Justification:

The purpose of this symposium is to present recent work aimed at improving the design and power analyses of impact studies of STEM interventions. It is often the case that these impact studies utilize a cluster randomized trial (CRT) in which students are nested within teachers nested within schools. The outcomes of interest may be at the student level, such as science achievement, or at the teacher level, such as teacher pedagogical content knowledge or content knowledge. In order for these CRTs to yield rigorous evidence of whether an intervention has a treatment effect at the student and/or teacher level, it must be designed with adequate power to detect a meaningful effect at either/both levels. The first paper in this symposium presents effect sizes from a meta-analysis of experimental and quasi-experimental studies of interventions with preservice and inservice science teachers, in the U.S. and internationally. The second paper provides results from secondary analyses with 8 existing educational datasets that have at least one teacher outcome in math or science and reports estimates of the intraclass correlation (ICC), and the percent of variance explained by covariate(s) ($R^2$). The empirically estimated design parameters (effect sizes, ICCs, $R^2$s) from these two papers will help researchers planning CRTs of STEM interventions conduct more accurate power analyses for teacher-level outcomes, which is critical given that lack of empirical estimates of design parameters for teacher-level outcomes. The third paper considers the design of CRTs that seek to impact both teacher outcomes and student outcomes in one study, such as an impact study that evaluates a STEM professional development program. This paper integrates the findings from papers 1 and 2 as well as recent literature on design parameters for student outcomes to examine the alignment of the power analyses when a study seeks to examine the impact of a STEM intervention on both teacher and student effects.
Title: Investigating Science Teacher Effect Sizes for A Priori Power Analyses

Presenting Author: Kowalski, S. M., BSCS


Purpose
As part of the larger project to examine a priori power analysis parameters for teacher outcomes from studies of science education interventions, we conducted a meta-analysis of experimental and quasi-experimental studies of interventions with preservice and inservice science teachers, in the U.S. and internationally. The results will inform study designs for primary researchers.

Research Question:
What effect sizes exist in the research literature on education interventions for science teachers and how do those effect sizes vary within and across designs or intervention characteristics?

Methods

Eligibility
All eligible studies were
- Published in 2001 or later
- Studied the effect of an intervention on preservice or inservice science teachers of primary or secondary students in U.S. and international contexts
- Published in English
- Included outcome measures for preservice or inservice science teachers
- Included baseline equivalence measures

We selected 2001 because it coincides with the passage of the No Child Left Behind act that played an important role in transforming the types of research conducted in educational settings in the U.S.. We wished to identify studies designed to meet that call for causal effects research. We included only studies published in English as we lacked capacity for translation.

Identifying Studies
We identified studies by searching education databases (e.g., ERIC, Dissertation Abstracts), web-based databases (Google Scholar, Yahoo), general education, policy journals, evaluation journals, science education journals, conference programs, research reports from IES and National Science Foundation funded studies, reports from national databases such as the WWC and the Best Evidence Encyclopedia, white papers, and reports from educational publishers’ websites.

Screening
We double-screened all abstracts and the PI resolved conflicts. We identified 250 potentially eligible studies based on abstract screening. We identified 164 unique studies for coding after screening full texts.
Coding Procedures
Using a database, we coded information related to research design, sample size and characteristics, treatment and comparison group characteristics, the nature of the intervention and of dependent variables, and effect size data. We calculated standardized mean difference effect sizes. Studies reported findings in many formats. We converted all statistics to d-family effect size indices (see Lipsey & Wilson, 2001).

Interrater Agreement
Coders jointly coded studies and met to discuss codes. During discussions, coders resolved any inconsistencies. They continued to jointly code one study at a time until they reached complete agreement (we initially aimed for 90% agreement, but surpassed that goal by the final two jointly coded studies). The remainder of the studies were divided among the coders.

Statistical Independence
When two studies included overlapping participant samples, we linked the two studies with a single Study ID and used all linked manuscripts to code the single study.

We coded all effect sizes for each study. We accounted for the lack of statistical independence of effect sizes (multiple effects within a single study) using Robust Variance Estimation (RVE) meta-regression (Hedges, Tipton, & Johnson, 2010).

Analysis
Statistical adjustments. We Winsorized sample sizes and effect sizes to ensure that extremely large values did not overly influence regression coefficients. We also adjusted sample sizes for studies that had cluster-level assignment (Higgins et al., 2011). Discrete predictor variables account for study design, sample characteristics, and intervention characteristics. We grand mean centered all variables. We gave more weight to studies whose effect size estimates have greater precision, where precision is primarily a function of study sample size. However, in our approach to weighting, we include a within-study as well as a between-study component to the variance. The within-study component is the average variance across effect sizes within the study, and the between-study component is calculated following Hedges, Tipton, and Johnson (2010).

Analytic model. Our RVE meta-regression equation is:

\[ ES_{ij} = \beta_0 + \beta_1 RCT + \beta_2 Pub + \sum_{k=3}^{6} \beta_k Sci_{ij} + \sum_{m=7}^{8} \beta_1 DV_{ij} \sum_{l=9}^{13} \beta_m TxComp_{ij} \\
+ \beta_{14} AltTx + \beta_{15} TxDiff + \beta_{16} AssesDev + \beta_{17} DurHrs + e_{ij} \]

Table 1 defines the variables.

Results
We coded 789 posttest effect sizes across the 164 studies. Table 1 summarizes the results of the regression using RVE (using the program Robumeta in R; Fisher, Tipton, & Hou, 2017). In general, we found a grand mean effect size across interventions of 0.42. Randomized studies tend to have lower effect sizes than quasi-experiments, biology interventions tend to have larger effect sizes than those in other science disciplines, effects tend to be larger when the outcome measured is a practice measure, studies that compare one treatment to another alternative
treatment tend to have smaller effects than those that compare a treatment to either business-as-usual or no treatment, and interventions that include curriculum materials for a teacher to use also tend to produce larger effect sizes than interventions that do not include curriculum as a treatment component. We are currently developing a calculator (R-based shiny app) that will allow users to estimate an a priori effect size for a planned intervention. The estimated effect sizes will allow authors to better estimate the number of participants needed in their studies when little other information about an effect size from prior research is available. One surprising finding is that publication status (peer reviewed or not) did not substantially impact the effect size ($\beta = -.08$). That said, our funnel plot (Figure 1) is not symmetrical indicating that we may not be adequately capturing publication bias in our Pub regression coefficient. Studies with low effect sizes may never have been shared in any publicly available format, peer reviewed or otherwise.

Conclusions
We found that effect sizes of teacher outcomes in science education intervention studies are highly variable. In some ways, the variation is expected. One might expect that randomized studies have smaller effect sizes than quasi-experiments. On the other hand, when controlling for randomization and other variables, publication status was not highly correlated with effect size. This finding contradicts our own previous work (Authors, 2017) and that of others (Polanin, Tanner-Smith, & Henessy, 2016). We also note that none of the coefficients are statistically significant predictors of effect sizes. Thus, the coefficients pertain specifically to our sample of studies. But given the wide-ranging nature of our search and the inclusion of such a wide array of experiments and quasi-experiments, we assert that the coefficients can provide valuable information to science education researchers as they design their own studies.
Table 1. Predictor variables in meta-regression model and their meaning.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning (1 = yes; 0 = no)</th>
<th>Reference group</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCT</td>
<td>Was the sample randomly assigned to treatment or comparison?</td>
<td>NA</td>
</tr>
<tr>
<td>Pub</td>
<td>Was the study published in a peer-reviewed journal?</td>
<td>NA</td>
</tr>
<tr>
<td>Chem</td>
<td>Was the science discipline chemistry?</td>
<td>biology</td>
</tr>
<tr>
<td>Phys</td>
<td>Was the science discipline physics?</td>
<td></td>
</tr>
<tr>
<td>PhysSci</td>
<td>Was the science discipline physical science?</td>
<td></td>
</tr>
<tr>
<td>EaMul</td>
<td>Was the science discipline Earth/Space or Multidisciplinary?</td>
<td></td>
</tr>
<tr>
<td>ContDV</td>
<td>Was the dependent variable a knowledge outcome?</td>
<td>attitudes</td>
</tr>
<tr>
<td>PracDV</td>
<td>Was the dependent variable a classroom practice outcome?</td>
<td></td>
</tr>
<tr>
<td>CurTx</td>
<td>Did the intervention include a curriculum component?</td>
<td>NA</td>
</tr>
<tr>
<td>SciTx</td>
<td>Did the intervention include the opportunity to learn science concepts?</td>
<td>NA</td>
</tr>
<tr>
<td>MentTx</td>
<td>Did the intervention include a mentorship component?</td>
<td>NA</td>
</tr>
<tr>
<td>AdMtITx</td>
<td>Did the intervention include additional materials (not curriculum)?</td>
<td>NA</td>
</tr>
<tr>
<td>CollTx</td>
<td>Did the intervention included the opportunity for teachers to collaborate?</td>
<td>NA</td>
</tr>
<tr>
<td>AltTx</td>
<td>Did the comparison group receive a focused treatment (rather than BaU)?</td>
<td>NA</td>
</tr>
<tr>
<td>TxDif</td>
<td>How many components did the treatment have that the comparison did not?</td>
<td>NA</td>
</tr>
<tr>
<td>AssessDev</td>
<td>Was the author the assessment developer?</td>
<td>NA</td>
</tr>
<tr>
<td>DurHrs</td>
<td>What was the duration of the study in hours?</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table 2. Regression estimates from Robust Variance Estimation (RVE) meta-regression.

<table>
<thead>
<tr>
<th>β Coeff</th>
<th>Variable Name</th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>dfs</th>
<th>p-value</th>
<th>95% CI L</th>
<th>95% CI U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Intercept</td>
<td>0.41678</td>
<td>0.17091</td>
<td>2.439</td>
<td>17.89</td>
<td>0.0254</td>
<td>0.05756</td>
<td>0.776</td>
</tr>
<tr>
<td>1</td>
<td>RCT</td>
<td>-0.32095</td>
<td>0.22741</td>
<td>-1.411</td>
<td>19.89</td>
<td>0.1736</td>
<td>-0.79547</td>
<td>0.1536</td>
</tr>
<tr>
<td>2</td>
<td>Pub</td>
<td>-0.08982</td>
<td>0.30095</td>
<td>-0.298</td>
<td>17.06</td>
<td>0.769</td>
<td>-0.7246</td>
<td>0.545</td>
</tr>
<tr>
<td>3</td>
<td>Chem</td>
<td>-0.52391</td>
<td>0.43765</td>
<td>-1.197</td>
<td>15.47</td>
<td>0.2493</td>
<td>-1.45426</td>
<td>0.4064</td>
</tr>
<tr>
<td>4</td>
<td>Phys</td>
<td>-0.5077</td>
<td>0.41712</td>
<td>-1.217</td>
<td>15.8</td>
<td>0.2414</td>
<td>-1.39285</td>
<td>0.3775</td>
</tr>
<tr>
<td>5</td>
<td>PhySci</td>
<td>-0.83323</td>
<td>0.61133</td>
<td>-1.363</td>
<td>4.71</td>
<td>0.2344</td>
<td>-2.43434</td>
<td>0.7679</td>
</tr>
<tr>
<td>6</td>
<td>EaMul</td>
<td>-0.36742</td>
<td>0.43449</td>
<td>-0.846</td>
<td>16.16</td>
<td>0.4101</td>
<td>-1.28776</td>
<td>0.5529</td>
</tr>
<tr>
<td>7</td>
<td>ContDV</td>
<td>0.12558</td>
<td>0.24075</td>
<td>0.522</td>
<td>18.85</td>
<td>0.608</td>
<td>-0.37858</td>
<td>0.6297</td>
</tr>
<tr>
<td>8</td>
<td>PracDV</td>
<td>0.47867</td>
<td>0.36555</td>
<td>1.309</td>
<td>12.12</td>
<td>0.2146</td>
<td>-0.31687</td>
<td>1.2742</td>
</tr>
<tr>
<td>9</td>
<td>CurTx</td>
<td>0.27147</td>
<td>0.50537</td>
<td>0.537</td>
<td>7.43</td>
<td>0.6069</td>
<td>-0.90964</td>
<td>1.4526</td>
</tr>
<tr>
<td>10</td>
<td>SciTx</td>
<td>0.17696</td>
<td>0.35721</td>
<td>0.495</td>
<td>9.65</td>
<td>0.6314</td>
<td>-0.62294</td>
<td>0.9769</td>
</tr>
<tr>
<td>11</td>
<td>MentTx</td>
<td>-0.50857</td>
<td>0.30678</td>
<td>-1.658</td>
<td>10.33</td>
<td>0.1274</td>
<td>-1.18921</td>
<td>0.1721</td>
</tr>
<tr>
<td>12</td>
<td>AdMtlTx</td>
<td>0.15388</td>
<td>0.28809</td>
<td>0.534</td>
<td>11.95</td>
<td>0.6031</td>
<td>-0.47412</td>
<td>0.7819</td>
</tr>
<tr>
<td>13</td>
<td>CollTx</td>
<td>-0.0906</td>
<td>0.21077</td>
<td>-0.43</td>
<td>18.16</td>
<td>0.6724</td>
<td>-0.53313</td>
<td>0.3519</td>
</tr>
<tr>
<td>14</td>
<td>AltTx</td>
<td>-0.34305</td>
<td>0.28009</td>
<td>-1.225</td>
<td>9.83</td>
<td>0.2492</td>
<td>-0.96857</td>
<td>0.2825</td>
</tr>
<tr>
<td>15</td>
<td>TxDif</td>
<td>-0.01724</td>
<td>0.10345</td>
<td>-0.167</td>
<td>18.76</td>
<td>0.8694</td>
<td>-0.23396</td>
<td>0.1995</td>
</tr>
<tr>
<td>16</td>
<td>AssessDev</td>
<td>0.18133</td>
<td>0.25423</td>
<td>0.713</td>
<td>23.56</td>
<td>0.4827</td>
<td>-0.34389</td>
<td>0.7066</td>
</tr>
<tr>
<td>17</td>
<td>DurHrs</td>
<td>0.00248</td>
<td>0.00431</td>
<td>0.575</td>
<td>13.82</td>
<td>0.5746</td>
<td>-0.00677</td>
<td>0.0117</td>
</tr>
</tbody>
</table>
Figure 1. Funnel plot of effect sizes and sample sizes (shows some bias in favor of larger effect size studies).
References

Studies WITHOUT asterisks were included in the meta-analysis. When multiple research reports refer to the same study sample, we coded the multiple reports as a single study.


Aydin, Serap Oz; Sahin, Selin; Sicaker, Aysun; Laborda, JC Ozdamli, F Maasoglu, Y. (2014). The effect of protein synthesis game in the class on the students' understanding of protein synthesis subject. Paper presented at the meeting of the.


Batiza, Ann Finney; Gruhl, Mary; Zhang, Bo; Harrington, Tom; Roberts, Marisa; LaFlamme, Donna; Haasch, Mary Anne; Knopp, Jonathan; Vogt, Gina; Goodsell, David; Hagedorn, Eric; Marcey, David; Hoelzer, Mark; Nelson, Dave (2013). The Effects of the SUN Project on Teacher Knowledge and Self-Efficacy regarding Biological Energy Transfer Are Significant and Long-Lasting: Results of a Randomized Controlled Trial CBE - Life Sciences Education, 12, 287-305.


Benli, Esra; Sarikaya, Mustafa; Baskan, GA; Ozdamli, F; Kanbul, S; Ozcan, D. (2012). The investigation of the effect of problem based learning to the academic achievement and the permanence of knowledge of prospective science teacher: the problem of the boiler stone. Paper presented at the meeting of the.


Bradshaw, T. J., Cunningham, J. D., Osborn, J. L., & Bradley, K. D. (2012). Impact of Newton’s Universe Materials on Middle School Teacher and Student Understanding of Physical


Cavalluzzo, Linda; Barrow, Lisa; Henderson, Stephen (2014).


Celikler, Dilek; Aksan, Zeynep; Yalin, HI Adiloglu, F Boz, H Karatas, S Ozdamli, F. (2011). The effect of computer assisted instruction in teaching ionic compounds on pre-service elementary science teachers’ academic achievement and permanent learning. Paper presented at the meeting of the , .


Diamond, Brandon S.; Maerten-Rivera, Jaime; Rohrer, Rose Elizabeth; Lee, Okhee (2014). Effectiveness of a curricular and professional development intervention at improving elementary teachers' science content knowledge and student achievement outcomes: Year 1 results Journal of Research in Science Teaching, 51(5), 635-658.


Enderle, Patrick; Dentzau, Michael; Roseler, Katrina; Southerl, Sherry; Granger, Ellen; Hughes, Roxanne; Golden, Barry; Saka, Yavuz (2014). Examining the Influence of RETs on Science Teacher Beliefs and Practice Science Education, 98, 1077-1108.


Glogger-Frey, Inga; Fleischer, Corinna; Grueny, Lisa; Kappich, Julian; Renkl, Alexander (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction Learning and Instruction, 39, 72-87.

Goldenberg, Lauren B.; Culp, Katherine McMillan; Clements, Margaret; Pasquale, Marian;erson, Alice (2014). Online Professional Development for High-School Biology Teachers: Effects on Teachers’ and Students’ Knowledge Journal of Technology and Teacher Education, 22, 287-309.


Hrepic, Zdeslav;Reed, RH Berque, DA Prey, JC (2009). Deploying Tablet PCs and DyKnow Software in Teaching an Inquiry-Based Physical Science Course Unpublished manuscript . .

Hrepic, Zdeslav (2008). Deploying Tablet PCs and DyKnow software in teaching inquiry-based physical science course In (Ed.), The Impact of Tablet PCs and Pen-based Technology on Education: Evidence and Outcomes (pp. 55-62). : Purdue University Press.


Karpudewan, Mageswary; Ismail, Zurida; Roth, Wolff-Michael (2012). Ensuring sustainability of tomorrow through green chemistry integrated with sustainable development concepts (SDCs) Chemistry Education Research and Practice, 13, 120-127.

Karpudewan, Mageswary; Ismail, Zurida; Roth, Wolff-Michael (2012). Promoting pro-environmental attitudes and reported behaviors of Malaysian pre-service teachers using green chemistry experiments Environmental Education Research, 18(3), 375-389.


Nugent, Gwen; Tol, Michael D.; Levy, Richard; Kunz, Gina; Harwood, David; Green, Denise; Kitts, Kathy (2012). The Impact of an Inquiry-Based Geoscience Field Course on Pre-Service Teachers Journal of Science Teacher Education, 23, 503-529.


Osmundson, Ellen; Dai, Yunyun; Herman, Joan (2011). Year 3 ASK/FOSS Efficacy Study. CRESST Report 782.


Qing, Zhou; Jing, Guo; Yazhuan, Liu; Ting, Wang; Junping, Ma (2010). Promoting preservice teachers' critical thinking Disposition by inquiry-based chemical experiment Procedia Social and Behavioral Sciences, 9, 1429-1436.


Sampson, Victor ;Enderle, Patrick ;Grooms, Jonathon ;Witte, Shelbie  (2013). Writing to learn by learning to write during the school science laboratory: Helping middle and high school students develop argumentative writing skills as they learn core ideas Science Education, 97(5), 643-670.


Selcuk , Gamze Sezgin;Caliskan , Serap;Uzunboylu , H. (2010). A small-scale study comparing the impacts of problem-based learning and traditional methods on student satisfaction in the introductory physics course Innovation and Creativity in Education, 2, 809-813.


Sendur , Gulten;Toprak , Mustafa (2013). The role of conceptual change texts to improve students' understanding of alkenes Chemistry Education Research and Practice, 14, 431-449.
Senocak, Erdal; Taskesenligil, Yavuz; Sozbilir, Mustafa (2007). A Study on Teaching Gases to Prospective Primary Science Teachers through Problem-Based Learning Research in Science Education, 37, 279-290.


Stoddart, Trish; Bravo, Marco; Mosqueda, Eduardo; Solis, Jorge (2013). Restructuring Pre-Service Teacher Education to Respond to Increasing Student Diversity Research in Higher Education Journal, 19, .


Tasoglu, Aslihan Kartal; Bakac, Mustafa; Uzunboylu, H. (2010). The effects of problem based learning and traditional teaching methods on students' academic achievements, conceptual developments and scientific process skills according to their graduated high school types Innovation and Creativity in Education, 2, 2409-2413.


Thompson, R. (2003). The effect of a laboratory-based, in-context, constructivist teaching approach on preservice teachers' science knowledge and teaching efficacy (Doctoral
Tosun, Cemal; Taskesenligil, Yavuz (2013). The Effect of Problem-Based Learning on Undergraduate Students' Learning about Solutions and Their Physical Properties and Scientific Processing Skills. Chemistry Education Research and Practice, 14, 36-50.


Tuysuz, Cengiz; Demirel, Ozan Emre; Yildirim, Bilal; Odabasi, HF (2013). Investigating the effects of argumentation, problem and laboratory based instruction approaches on pre-service teachers' achievement concerning the concept of 'acid and base'. Paper presented at the meeting of the.


Walker, M., Recker, Mimi; Ye, Lei; Robertshaw, M. Brooke; Sellers, Linda; Leary, Heather (2012). Comparing technology-related teacher professional development designs: a
multilevel study of teacher and student impacts Etr&D-Educational Technology Research and Development, 60, 421-444.


Title: Estimates of IntraClass Correlation and Outcome-Covariate Correlations for Teacher Outcomes in Evaluations of Math and Science Interventions

Presenting Author: Fatih Unlu, RAND

Authors: Fatih Unlu, RAND; Joe Taylor, BSCS; Jessaca Spybrook, WMU; Carl Westine, UNCC; and Brent Anderson, RAND

Background:

The past fifteen years have seen a dramatic increase in the use of cluster randomized trials (CRTs) to evaluate educational interventions. In a CRT, entire clusters (such as schools) are randomly assigned to the study conditions. In order to yield rigorous evidence on program effectiveness, a CRT must have a strong design, implementation, and analysis. This paper focuses on one of these components, the design of CRTs, and specifically improving the power analyses. The research on power analyses highlights the importance of using accurate, relevant, and reliable design parameters noting that the power analysis is only as good as the design parameters. This paper focuses on two specific design parameters: the percent of variance in the outcome measures between clusters, the intraclass correlation (ICC), and the percent of variance explained by covariate(s) (R^2).

Prior work on design parameter estimates suggests that there is substantial variation in these parameters across outcome domains, measures within a domain, subject matters, and grade-levels (e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2013; Phelps, Kelcey, Jones, & Liu, 2016; and Spybrook, Westine, & Taylor, 2016). While the literature covers student achievement quite well, only a few studies provide estimates for teacher outcomes such as use of research based instructional practices, content knowledge, and pedagogical content knowledge. The evidence is sparser for teacher outcomes in math and science than reading. This is a critical limitation for educational evaluations because teacher outcomes are at least as important as student outcomes. For example, the interventions in the early development stage may choose to initially focus on teacher-level outcomes rather than student-level outcomes because the former are more proximal to the intervention. Therefore, examining the former would give an early indication of the potential effects on the latter and may greatly inform and enhance the intervention development.

Purpose:

Education researchers designing studies to evaluate interventions with teacher outcomes in math and science often have to borrow design parameter estimates from other measures and subject—setting up a high likelihood for inappropriately powered studies, inconclusive results, and stagnation of knowledge in science and math education research. This paper addresses this limitation by conducting secondary analyses with 8 existing educational datasets that have at least one teacher outcome in math or science and reporting estimates of ICC and R^2 values. One of these datasets is from the Measures of Effective Teaching (MET) study and the other seven are from relatively large and federally funded studies. We have selected these data sets because they provide us the unique opportunity to track outcomes from both national and regional samples of individual teachers over time (i.e., for baseline-outcome R^2 analysis), and contain a grouping variable for estimating cluster effects (i.e., for ICCs). Further, these data sets include data on other teacher characteristics (e.g., content knowledge, years of experience) for use in

1 One study is funded by NSF and the others are funded by the Institute of Education Sciences.
other R² analyses. The combined samples of these datasets include over 6,000 teachers in more than 1,000 schools.

We currently have access to all datasets and have completed preliminary analyses of two. We will complete the analyses of the remaining datasets by the time of the conference and include the corresponding results in our presentation should our proposal be accepted.

Method:

The design parameters are estimated using 2-level hierarchical linear models (HLMs) with teachers nested within schools.² First, we estimate an unconditional HLM specified as:

\[
\text{Level 1 (teacher): } Y_{ij} = \alpha_{0j} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2) \quad [1]
\]

\[
\text{Level 2 (school): } \alpha_{0j} = \gamma_0 + \lambda_{0j}, \quad \lambda_{0j} \sim N(0, \nu^2) \quad [2]
\]

where \(Y_{ijk}\) is the outcome (e.g., instructional practices) for teacher \(i\) in school \(j\); \(e_{ij}\) is the error associated with teachers and \(\sigma^2\) is the teacher-level variance; and \(\lambda_{0j}\) is the error associated with schools and \(\nu^2\) is the school-level variance. The ICC coefficient is given by:

\[
ICC_{L2} = \frac{\nu^2}{\sigma^2 + \nu^2} \quad [3]
\]

To calculate the variance explained by covariates, the unconditional model is modified to include teacher and school-level covariates. We examine three covariate sets: (1) baseline version of the outcome (i.e., pretest) only; (2) teacher and school demographic characteristics only; and (3) full covariate set including the pretest and all other teacher and school-level baseline characteristics. We define two R² terms that capture the proportion of outcome variance at the teacher and school level explained by the covariates:

\[
R^2_{L1} = \frac{\sigma^2 - \sigma^2_{|X}}{\sigma^2} \quad \text{and} \quad R^2_{L2} = \frac{\nu^2 - \nu^2_{|X,W}}{\nu^2} \quad [4]
\]

where \(X\) and \(W\) are vectors containing teacher and school-level covariates and \(\sigma^2_{|X}\) and \(\nu^2_{|X,W}\) are conditional versions of the teacher and school-level variance terms. We also report standard errors of the ICC and R² estimates using formulae provided by Siddiqui et al. (1996) and Olkin and Finn (1995).

---

² We also estimate 3-level models – teachers nested within schools nested within districts – with datasets that have sufficiently large number of schools per district.
Results:

Table 1 shows the ICC and R2 estimates from the preliminary analyses. The estimates vary substantially across studies and across outcomes within each study. This variation has implications for statistical power, minimum detectable effect sizes, and sample size requirements. Specifically, the ICC values in the first dataset are considerably larger than those in the second dataset; therefore, using the former would lead to larger sample size requirements than the latter. One factor underlying this variation could be the different levels of heterogeneity of teachers and schools across the two datasets.

The observed variation across datasets highlights the importance of calculating design parameters with samples from different settings and locations and providing contextual and descriptive information about the samples so that the users of the estimated design parameters could choose the values for their designs accordingly. The completed version of the paper will provide this information, examine factors that may explain the variation in the design parameter estimates, and assess the generalizability of the parameter estimates using the approach described in Jacob, Zhu, and Bloom (2010).
Table 1. Estimated Values of ICC and R² Coefficients

<table>
<thead>
<tr>
<th>Measure</th>
<th>Subject</th>
<th>Dataset</th>
<th>ICC</th>
<th>Level 1 (Teacher) R²</th>
<th>Level 2 (School) R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Estimate</td>
<td>Std Error</td>
</tr>
<tr>
<td>Instructional Practices</td>
<td>Science</td>
<td>Science Teachers</td>
<td>0.57</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Content Knowledge</td>
<td>Science</td>
<td>Learning through Lesson Analysis1</td>
<td>0.3</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>Pedagogical Content Knowledge</td>
<td>Science</td>
<td>Lesson Analysis1</td>
<td>0.43</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Mathematical Quality of Instruction</td>
<td>Math</td>
<td>Measures of Effective Teaching</td>
<td>0.03</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>Quality of Science Teaching</td>
<td>Science</td>
<td>Learning through Lesson Analysis1</td>
<td>0.11</td>
<td>0.32</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

Notes: ¹ For all outcomes, the sample includes 133 teachers in 77 schools. R² values are from conditional models that control for baseline measure of the outcome. ² The sample includes 767 teachers in 225 schools for the math outcome and 45 teachers in 33 schools for the science outcome. R² values are from conditional models that control for baseline measure of the outcome.
References


Title: An Investigation of Design and Statistical Power for Planning Cluster Randomized Trials with Student and Teacher Outcomes

Presenting Author: Qi Zhang, Western Michigan University

Authors: Qi Zhang and Jessaca Spybrook, Western Michigan University; Fatih Unlu, RAND

Background: Cluster randomized trials (CRTs) are commonly conducted to assess the efficacy of educational interventions, which are often implemented at the school level. CRTs with schools as the natural unit of random assignment often have teachers nested within schools and students nested within teachers. Further, it is often the case that the outcomes of interest are at both the student- and the teacher-level. For instance, CRTs designed to evaluate the efficacy of teacher professional development (PD) programs are interested in determining the effect of PD programs on teacher content knowledge and teacher practice, as well as their impact on student achievement. Since design parameters used to calculate the statistical power, such as the intra-class correlation coefficients (ICCs) and the outcome-covariate correlations (R² coefficients) tend to be different for the teacher and the student outcomes, a single power analysis is generally not sufficient to assess whether the study is adequately powered to detect effects for teachers and students. Therefore, researchers need to conduct two power analyses—one for teacher outcomes and one for student outcomes—and determine sample size requirements according to both analyses, as a design that is adequately powered to detect a meaningful impact at the student level is not necessarily powered to detect a meaningful impact at the teacher level and vice versa.

Purpose: This paper examines design considerations for studies that seek to evaluate the effectiveness of educational interventions for both teachers and students within one study. Specifically, this paper incorporates new empirical work presented in this symposium on design parameters for planning CRTs with student and teacher outcomes to provide insights for CRT designs of K-12 science educational interventions. The goal of this paper is to estimate the statistical power of these studies, examine the alignment of the power analyses when a study seeks to examine both teacher and student effects, and suggest considerations that can be incorporated into future planning to maximize the efficiency of the study design.

Method: We calculated statistical power, represented by the minimum detectable effect size (MDES) for CRTs that examine interventions that seek to improve student science achievement, as well improvements in science teacher content knowledge, teacher pedagogical content knowledge, and teacher practice. We assumed school as the unit of random assignment and allowed the total number of schools to vary from 25 to 65 schools. For the teacher level outcomes, we assume 5 teachers nested within each school, which is common within a particular grade in elementary schools. Assuming 25 students nested within each teacher, we set the number of students per school to 125 for the design to assess student outcomes. We compared MDES for studies with student and teacher outcomes based on best estimates of design parameters.

The power calculations for student outcomes are based on a 2-level Hierarchical Linear Model (HLM) that nested students within schools to estimate impacts. We ignored the teacher-level for this calculation, which is not expected to influence the calculated MDES (Zu, Jacob, Bloom, & Xu, 2012). Table 1 shows the design parameters used for this calculation, which are based on those reported by Westine, Spybrook, & Taylor (2014) for grade 5-11 science achievement outcomes. For the ICC and the school-level R², we used an upper and lower bound. We used a single value for the individual-level R². Different
combinations of these parameter values lead to a range for the MDES estimates, which are discussed below.

The MDES calculations for teacher outcomes are based on a 2-level HLM that nests teachers within schools. The design parameters for these calculations were based on plausible values from the empirical analyses presented in paper 2 of this symposium for teacher practice and content knowledge outcomes.

Table 1. Empirical estimates used for the MDES calculations*.

<table>
<thead>
<tr>
<th></th>
<th>School-level ICC</th>
<th>School-level (Level-2) $R^2$</th>
<th>Individual-level (Level-1) $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Outcome</td>
<td>0.20, 0.26</td>
<td>0.50, 0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>Teacher Outcome</td>
<td>0.15, 0.30</td>
<td>0.10, 0.20</td>
<td>0.25</td>
</tr>
</tbody>
</table>

*Calculations were based on these additional assumptions: two-tailed test, alpha = 0.05, equal allocation at all levels.

Equation (1) below shows the formula for calculating MDES for a 2-level CRT and we operationalized the formula in the program PowerUp! (Dong & Maynard, 2013).

$$MDES_{2LCRT} = M_{j-3} \sqrt{\frac{\rho(1-R_{l2}^2)}{P(1-P)f}} + \frac{(1-\rho)(1-R_{l1}^2)}{P(1-P)f}$$

Equation (1)

Where $n$ is the number of individuals (teachers or students) per cluster and $J$ is the number of clusters. $M$ is the multiplier for two-tailed test with $J - 3$ degrees of freedom. $\rho$ is the ICC, which is the proportion of variance in outcome that is between clusters. $R_{l1}^2$ and $R_{l2}^2$ are the proportion of variance explained by level-1 and level-2 covariates, respectively. $P$ is the proportion of level-2 units randomized to treatment.

Results: Figure 1 shows the MDES calculated based on student achievement outcome (dashed line) and teacher outcomes (solid line). As the number of schools approach 45, the range of MDES for student outcomes (0.18–0.31) is similar to the effect sizes for educational interventions noted by Hill, Bloom, Black, and Lipsey (2008) (0.14-0.24). For the same number of schools, note that the MDESs for teacher content knowledge outcomes was higher (0.43-0.52), due to the small number of teachers per school, the larger value of ICC, and the smaller school-level $R^2$. However, the range for teacher content knowledge outcome was consistent with the median effect size (0.52) observed for science teacher interventions in paper 1 of this symposium. As mentioned in paper 2 of this symposium, the final version of this paper will incorporate results from our analysis of a large number of additional existing datasets. Further, we will optimize our calculations with a larger set of parameters, including varying number of teachers and baseline data collection activities.
Figure 1. Calculated MDES based student achievement outcome (dashed line) and teacher content knowledge outcome (solid line), with varying number of schools.

**Conclusion:** Our result suggested studies that include at least 45 schools, 5 teachers per school, and 25 students per teacher may be able to detect a meaningful effect of the intervention for both students and teachers. This was possible because the larger effect size for teachers observed in the meta-analysis study (paper 1 of the symposium) compared to the effect size for students. Studies with less than 45 schools may be able to detect meaningful effects for students, but they may have some difficulties to achieve the same for teachers. This study will further extrapolate the results of power calculations for other teacher outcomes, as well as the implication of these results in planning CRTs that examine intervention effects both student and teacher.
References


