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, Jessaca Spybrook, Western Michigan University; Fatih Unlu, RAND

Introduction

There has been a strong push for research on interventions of science education in the U.S. in the last decade. The number of studies conducted to examine the efficacy and effectiveness of math and literacy educational interventions greatly exceeds the number of studies done on science interventions. For instance, there are 157 Randomized Controlled Trials (RCTs) on mathematics interventions and 249 RCTs on literacy interventions reviewed by the What Works Clearinghouse (WWC) that met standards either with or without reservation since its establishment in 2002. Only 13 RCTs on science interventions reviewed that met standards. This gap in the number of RCTs that met WWC standards is an indication that research, especially research that is grounded in strong methodology, is much needed in science education.

In the last decade, important federal policies have been established to support the increasing need to improve science education in K-12 settings and research in science education. The National Research Council, the National Science Teacher Association, and the American Association for the Advancement of Science introduced the Next Generation Science Standards (NGSS) in 2011, which is a multi-state effort to standardized science curriculum and science teaching in the U.S. (National Research Council, 2011). NGSS has gained a great amount of traction since its debut. As of 2017, 19 states have adopted this new standard into the Common Core standard and more states are expected to implement NGSS in the upcoming years. In parallel to the launch of the NGSS, The Institute of Educational Sciences (IES), the research branch of the U.S. Department of Education and the National Science Foundation joined forces in 2013 to create the Common Guideline for Educational Research and Development, with the goal of launching "cross-agency guidelines for improving the quality, coherence, and pace of knowledge development in science, technology, engineering, and mathematics (STEM) education" (IES & NSF, 2013). The guideline outlines four main types of research for knowledge generation: Foundation, Early-State or Exploratory, Design and Development, and Impact Research. For Impact Research, the guideline emphasizes that the study design should align with the WWC standards.

With the recent development of improving K-12 science curriculum and the increasing emphasis on impact and effectiveness studies of science interventions, it is expected that more impact studies will emerge. This is evident through the trend of federal funding within the last decade. The IES provides funding for research conducted in education based on four goals: Exploration, Development and Innovation, Efficacy and Replication, and Effectiveness. Efficacy and Replication or Goal-3 grants support studies that evaluate the impact of interventions. Development and Innovation or Goal-2 grants aim to support the development of innovative educational interventions that could produce beneficial impacts on students, which can be considered as the foundation for Goal-3 impact studies. In 2007 to 2017, 17 Goal-2 grants targeted science education and 12 grants targeted math education in the STEM Education program. This substantial number of Goal-2 grants funded in the last decades is an indication that more attention is placed on developing new science interventions, which could lead to more impact studies on these interventions.

A Priori Power Analysis

As the Common Guideline states, a key feature impact research, which includes Efficacy, Effectiveness, and Scale-up research, should address is the study design that the estimates of causal effect are based on. Specifically, a study design should contain elements that would produce strong causal validity, such that it would meet the WWC standards without reservation. An important feature in designing a rigorous RCTs is determining the sample size and statistical power associated with estimating the causal effect, which can be determined via a power analysis. It is within the researcher's interest to conduct a priori power analysis when designing an impact study. In this study, we are specifically interested in the power analysis for Cluster Randomized Trials (CRTs). Cluster randomized trials (CRTs) are commonly conducted to assess the efficacy of educational interventions, which are often implemented at the school level. For example, simple CRTs with schools as the natural unit of random assignment and student nested within schools are called 2-level CRTs. 3-level CRTs have randomization occurs the school level, with teacher nested within schools and students nested within teachers.

An a priori power analysis for a CRT is more complex than a power analysis for a single-level RCT. A researcher has to consider the percent of variation in the outcomes relative to the total variance that is between clusters, or the intra-class correlation (ICC). In a 2-level CRT with schools as the unit of randomization and students nested within schools, ICC describes the variation in student outcome (i.e. student science achievement) between schools. In this case, an ICC of 0.20 suggests that 20% variation of the student science achievement occurs between schools. The lower the variation is between schools or more homogenous the schools are in student outcome, the higher the statistical power, all other parameters held constant. A researcher may also want to consider including covariates that explain the variation in the outcome, which is represented in a percentage of variance explained (R^2 coefficient). For instance, student achievement pretest is commonly employed as a covariate for a power analysis, as it is often highly correlated with the achievement outcome of interest. In this case, an R^2 coefficient of 0.8 suggests that the covariate pretest explains 80% of the variance in the student outcome. When estimating power for CRTs, both individual-level and cluster-level covariate can be included to increase statistical power.

Statistical Power for Studies with Teacher and Student Outcomes

CRTs are often designed to detect the causal effects of interventions for both students and teachers. For instance, CRTs designed to evaluate the efficacy of teacher professional development (PD) programs may seek to determine 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 ICCs and R^2 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 because a design that is adequately powered to detect a meaningful impact at the student level may not necessarily be powered to detect a meaningful impact at the teacher level and vice versa.

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 on design parameters for planning CRTs with student and teacher outcomes to provide insights into 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 improve 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. This range is more similar to the median number of 20 clusters randomized for Goal-3 Efficacy and Replication studies funded by the IES in 2002-2004 and the median number of 52 clusters for studies funded in 2011-2013 found by Spybrook and colleagues (2013). For the teacher level outcomes, we assumed two scenarios. One with 5 teachers nested within each school, which is common in a particular grade in elementary school. Another scenario has 3 teachers nested within each school, which is more common in a particular grade in middle and high schools. Assuming 25 students nested within each teacher, we set the number of students per school to 75 or 125 depending on the number of teacher per school for the design to assess student outcomes. We compared the 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 R^2, we used a higher value and a lower value from the empirical estimates described by Westin, Spybrook, & Taylor (2014). We used a single value for the individual-level R^2. Different combinations of these parameter values lead to a range of the MDES estimates, which are discussed below.

The MDES calculations for teacher outcomes are also 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 for teacher practice and content knowledge outcomes (Unlu, Taylor, Spybrook, Westin, & Anderson, 2018) The two teacher outcomes have different range of ICC at the school level. We set the school-level R^2 range to be the same for both teacher content knowledge and practice outcomes. For individual-level R^2, we set the same single value for both teacher outcomes.

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

Outcome Measure	School-level ICC	School-level (Level-2) R^2	Individual-level (Level-1) R^2
Student Science Achievement	0.20, 0.26	0.50, 0.80	0.40
Teacher Content Knowledge	0.09, 0.21	0.10, 0.20	0.25
Teacher Practice	0.24, 0.36	0.10, 0.20	0.25

*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).

$$MDES_{2LCRT} = M_{J-3}\sqrt{\frac{\rho(1-R_{L2}^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_{L1}^2)}{P(1-P)Jn}} \qquad \text{Equation (1)}$$

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. ρ 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.

Result

Figure 1 shows the MDES calculated based on student achievement outcome (dashed line) and teacher outcomes (solid line). As the number of schools approached 40, the range of MDES for student outcomes (0.18–0.31) was 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 outcome (solid red line) was higher (0.42-0.51), due to the small number of teachers per school, the larger value of ICC, and the smaller school-level R^2. However, the range of MDES for teacher content knowledge outcome was consistent with the mean effect size of 0.44 observed for science teacher interventions found in a meta-analysis study of the effect of educational interventions for science teachers (Kowalski et al., 2018). The MDES for teacher practice outcome (solid blue line) was higher than the MDES for teacher content knowledge outcome, which is the result of larger values of ICC for teacher practice outcome. This indicated that 55 schools were necessary for MDES (0.43-0.50) for teacher practice outcome to contain the effect size of 0.44.

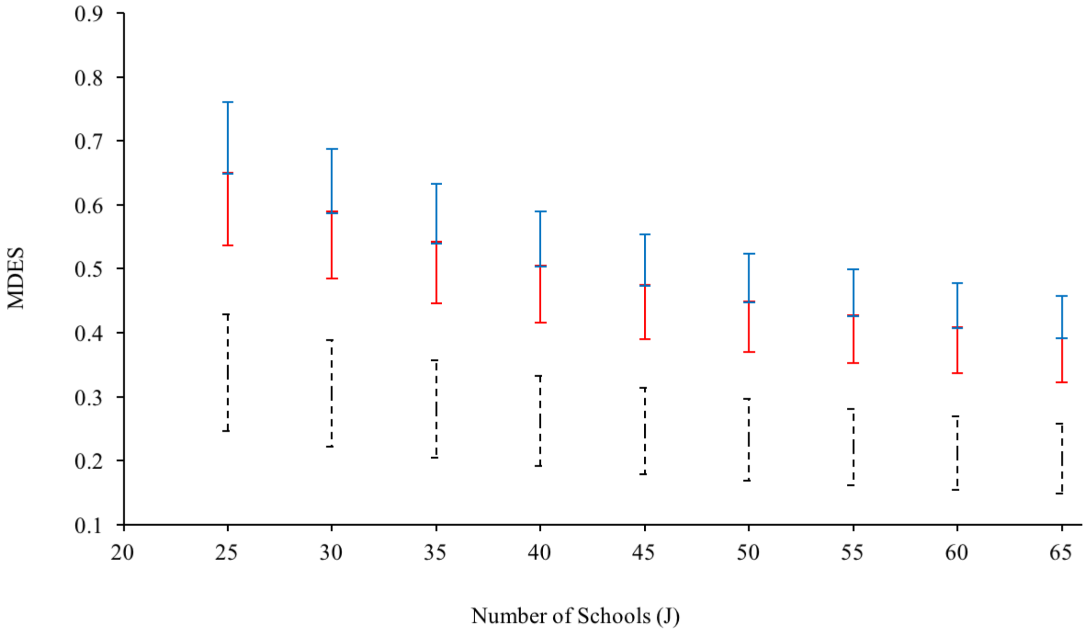

Figure 1. Calculated MDES based student achievement outcome (dashed line), teacher content knowledge outcome (solid red line), teacher practice outcome (solid blue line), with varying number of schools and 5 teachers per school.

We also calculated the MDES with 3 teachers per school and 25 students per teacher. The result of the calculation in comparison with the MDES for student science achievement was shown in in Figure 2 below. Note that the overall MDES was higher with the smaller number of teachers per school. With only 3 teachers per school, 50 schools were necessary to achieve the MDES of 0.44 – 0.50 for teacher content knowledge outcome, which contained the mean effect size of 0.44 for science interventions based on teacher outcomes. For teacher practice outcome, the MDES range of 0.44 – 0.49 was reached with 65 schools.

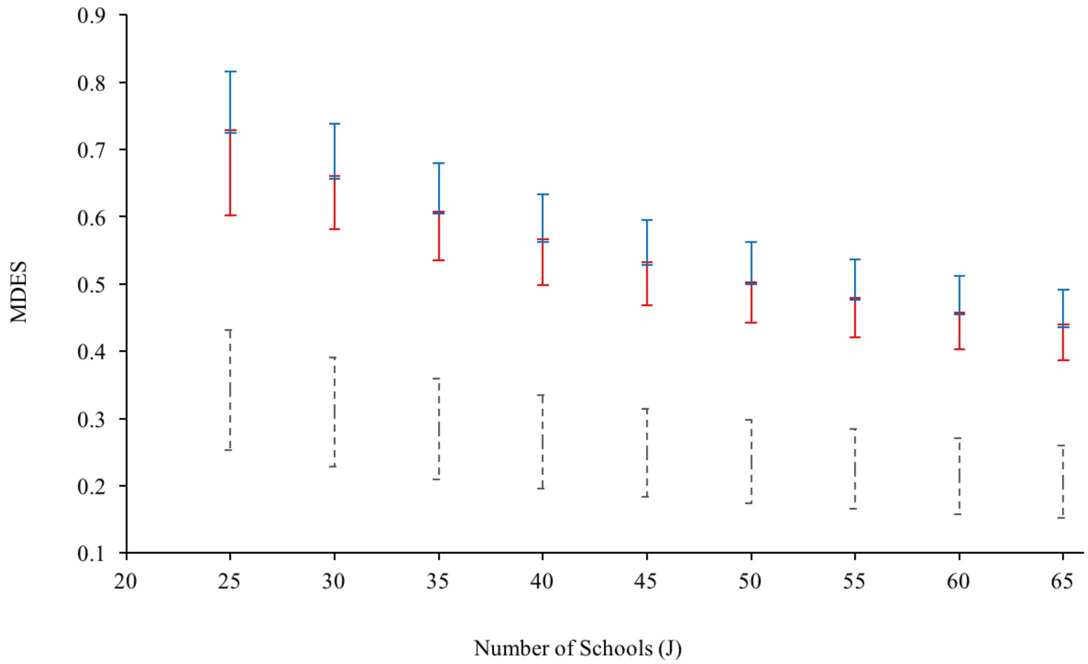

Figure 2. Calculated MDES based student achievement outcome (dashed line), teacher content knowledge outcome (solid red line), and teacher practice outcome (solid blue line) with varying number of schools and 3 teachers per school.

Discussion

The result of our study suggested some important design elements researchers need to consider when planning a study that aims to detect causal effects of the intervention on both teachers and students. The foremost important design elements, in regard to statistical power, is having sufficient sample sizes to optimally align the power to detect meaningful effect for both teacher and student outcomes. Our result suggested studies that include at least 40 schools, 5 teachers per school, and 25 students per teacher may be able to detect a meaningful effect when the outcome of interest are student science achievement and teacher content knowledge. This was possible because the larger effect size for teachers observed in the meta-analysis study (Kowalski et a., 2018) compared to the effect size for students. Studies with less than 40 schools may be able to detect meaningful effects for students, but they may have some difficulties to achieve the same for teachers. The number of schools required for a study planning to measure practice as the teacher outcome need at least 55 schools, 5 teachers per school, and 25 students per teacher. This larger number of schools is due to the greater variation in teacher practice occurs between schools, indicated by the relatively higher ICC values. When the number of teachers was limited to 3 per school, the number of schools necessary to adequately power a study that would detect a meaningful effect for teachers increased, while the sample sizes required to power a study for students remained the same. In this scenario, at least 50 schools were needed to adequately power a study based on teacher content knowledge outcome and 65 schools were needed based on teacher practice outcomes. These numbers were much higher than the 40 schools necessary to power a study based on student science achievement. If the number of schools in a study design is based on teacher outcomes with 3 teachers per school and 25 students per teacher, the study may be *over-powered* to detect the causal effect for students. Conversely, if the number of schools is based on student science achievement with 25 students per teacher or 125 students per school, then the study may be *under-powered* to detect the causal effect for teachers.

When aligning the study designs to detect meaningful effects for teachers and students, the result of our study also suggested that results may differ depending on the outcomes of interest. Our calculations indicated that a smaller number of schools was necessary to detect an effect for teachers based on content knowledge outcome measures than teacher practice outcomes. This discrepancy in the number of schools is partially due to the greater between-school variations in teacher practice outcomes. When deciding the teacher outcomes to include in a study, researchers need to estimate the amount of variation in that outcome within the sample. Further, researchers could consider including other variables that could explain some of the variations in the outcomes of interest. In this study, we estimated statistical power using pretest measures to help explain the variations in the teacher and the student outcomes. Other variables, such as student's eligibility in the free/reduced lunch program and years of teaching experience for teachers could be additional covariate to include to increase the precision of the power analysis. It is important to note that the difference in the MDES based on student and teacher outcomes is influenced by the difference in the R^2 coefficients associated with each types of outcomes at the school and individual level. The school-level R^2 value is especially important in the power analysis for a simple CRTs, where the units of random assignment are schools. For instance, the school-level R^2 coefficient for the student outcome range from 0.50 – 0.80 in this study, which is considerably higher than the corresponding R^2 coefficients of 0.10 – 0.20 for teacher outcomes. This difference in R^2 resulted in a higher MDES based on teacher outcomes. Therefore, a researcher needs to carefully select covariates that would best explain the variation in teacher outcomes within the context of the study.

The goal of understanding the variation in the outcomes of interest and incorporating variables that explain the variation in outcomes is to have a sample size that would optimally power the detection of causal effect for both teachers and students without having to either over-power or under-power a study. Ultimately, aligning the study design based on student and teacher outcomes ensures that the study is cost-effective and that it is sufficient to detect a meaningful effect of the intervention.

References

Dong, N**.**, & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.

Hill, C. J.; Bloom, H. S.; Black, A. R.; & Lipsey, M. (2008). Empirical Benchmark for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2(3), 172-177.

Institute of Educational Sciences & National Science Foundation. (2013). *Common Guidelines for Educational Research and Development.* Retrieved from https://ies.ed.gov/pdf/CommonGuidelines.pdf

National Research Council. (2011). A Framework for K-12 Science Education: Practices, Cross-cutting Concepts, and Core Ideas. Washington, DC: National Research Council.

Unlu, F., Taylor, J., Spybrook, J., Westin, C., & Anderson, B. (2018). *Estimates of IntraClass Correlation and Outcome-Covariate Correlations for Teacher Outcomes in Evaluations* *of Math and Science Interventions.* Paper presented at the Society of Research on Educational Effectiveness (SREE) Conference, Washington, D.C.

Kowalski, S., Taylor, J., Askinas, K., Anderson, D., Maddix, W., Wang, Q, & Zhang, Q. (2018). *Investigating Science Teacher Effect Sizes for A Priori Power Analyses.* Paper presented at the Society of Research on Educational Effectiveness (SREE) Conference, Washington, D.C.

Spybrook, J., Shi, R., & Kelcey, B. (2013). Progress in the Past Decade: An Examination of the Precision of Cluster Randomized Trials Funded by the U.S. Institute of Educational Sciences. *International Journal of Research & Method in Education*, 39(3), 255-267.

Westine, C. D.; Spybrook, J.; & Taylor, J. A. (2013). An Empirical Investigation of Variance Design Parameters for Planning Cluster-Randomized Trials of Science Achievement. *Evaluation Review*, 37(6), 490-519.

Zhu, P.; Jacob, R.; Bloom, H.; & Xu, Z. (2012). Designing and Analyzing Studies that Randomized Schools to Estimate Intervention Effects on Student Achievement Outcomes without Classroom-level Information. *Educational Evaluation and Policy Analysis*, 34(1), 45-68.