Educational Evaluation and Policy Analysis
September 2020, Vol. 42, No. 3, pp. 354–374
DOI: 10.3102/0162373720929018
Article reuse guidelines: sagepub.com/journals-permissions

Learning From Cluster Randomized Trials in Education: An Assessment of the Capacity of Studies to Determine *What Works*, *For Whom*, and *Under What Conditions*

Jessaca Spybrook
Qi Zhang
Western Michigan University
Ben Kelcey
University of Cincinnati
Nianbo Dong
The University of North Carolina at Chapel Hill

Over the past 15 years, we have seen an increase in the use of cluster randomized trials (CRTs) to test the efficacy of educational interventions. These studies are often designed with the goal of determining whether a program works, or answering the what works question. Recently, the goals of these studies expanded to include for whom and under what conditions an intervention is effective. In this study, we examine the capacity of a set of CRTs to provide rigorous evidence about for whom and under what conditions an intervention is effective. The findings suggest that studies are more likely to be designed with the capacity to detect potentially meaningful individual-level moderator effects, for example, gender, than cluster-level moderator effects, for example, school type.

Keywords: *educational policy, program evaluation, evaluation, experimental design, research methodology*

In the past 15 years, we have seen an increased emphasis on the use of randomized trials (RTs), particularly cluster randomized trials (CRTs), to test the efficacy of educational interventions (Spybrook et al., 2016). In a CRT, entire clusters, most often schools, are randomly assigned to condition. The most prominent funder of CRTs to assess the efficacy of educational interventions is the Institute of Education Sciences (IES), the research branch of the U.S. Department of Education. Since its inception in 2002, IES has played a leadership role in shaping education policy and practice around the use of RTs and CRTs to assess the efficacy of educational programs (Cook & Foray, 2007). As the leading funder of education research, IES has funded more than 250 efficacy studies, many of which are CRTs. Although there certainly are other federal funders of CRTs aside from IES, for example, the National Science Foundation (NSF) and the Office of Investment and Innovation (OII), IES has undoubtedly played a leadership role in the movement toward the use of CRTs to test educational interventions and as such is featured prominently in this article.

IES was established by the Education Science Reform Act of 2002. The mission of IES is to build a body of rigorous evidence to inform education policy and practice (http://ies.ed.gov). In the early years, this meant that IES focused on

answering the *what works* question. For example, *what math curriculum is most effective for improving math achievement? What reading professional development program is most effective for improving reading achievement?* IES prioritized efficacy studies that included RTs, most often CRTs, to answer this question. Although answering the *what works* question is still a primary goal of efficacy trials, over time we have seen the goals expand to questions about *for whom* and *under what conditions*. These types of questions are critical in that they help provide a more comprehensive picture about the types of students and schools that the programs are most impactful. For example, a new math program may be more effective for girls than boys or small schools rather than large schools. This is important information for school administrators as they consider whether or not an intervention will work in their schools and for their students. These questions are also central to the mission of IES to improve outcomes for all students.

Changes in the IES Request for Applications

As the goals of efficacy studies have evolved, so have the methodological expectations for the design of these studies. Although there are many important methodological components to the design of a rigorous efficacy study, we limit the scope of this article to one component, the statistical power to detect effects, specifically main effects and moderator effects. The main effect corresponds to the *what works* question, whereas the moderator effects correspond to the *for whom* and the *under what conditions* question. To document the changes in the methodological expectations around statistical power for main effects and moderator effects, we reviewed the IES requests for applications (RFA) from 2002 to 2017. Specifically, we focused on the guidelines for efficacy studies including the methodological requirements described in two relevant sections of the RFA, the section on statistical power and the section on the description of moderating variables. In the early RFAs, the language in the section on statistical power specified that a power analysis for the main effect of treatment was necessary. However, the details about what to include in the power analysis were limited. The study of moderating variables was encouraged in the section describing moderating variables. However, there was no mention of including a power analysis for moderator effects.

Over time, the requirements for the level of detail corresponding to the power analyses for the main effect of treatment were strengthened. For example, the RFA in the field year 2007 (Institute of Education Sciences, 2006) includes language stating that applicants must provide details related to the power analyses for the main effect of treatment and must justify the expected effect size. Furthermore, applicants planning a CRT should consider the total number of schools as well as the number of individuals per school and other relevant design parameters for CRTs. Spybrook and Raudenbush (2009) and Spybrook et al. (2016) examined the statistical power of studies to detect the main effect of treatment in IES-funded CRTs. The findings from these studies suggest that the precision of IES-funded studies has increased over time.

The requirements for the description of plans to assess moderating variables changed very little between 2002 and 2012. The RFA for the field year 2012 (Institute of Education Sciences, 2011) represents the first RFA to mention statistical power as it relates to moderating variables. In the section on the description of the moderating variables, the RFA stated,

> The Institute expects efficacy studies to examine relevant moderating factors. . . . The Institute recognizes that many efficacy studies are not powered to test the effects of a wide-range of moderators and so expects applicants to focus on a small set of well-justified ones.

Notably, in the section on the power analysis, the RFA did not explicitly ask for a power analysis for moderators.

The next key change related to methodological expectations around moderator effects occurred in the field year 2017 (Institute of Education Sciences, 2016). The RFA stated that the analysis of moderators is not required but rather that it makes for a stronger application in the section on the description of the moderators in the RFA. However, in the section on power, the RFA asks applicants to provide detailed power analyses for moderation, even if the moderator questions are considered exploratory. This represents a clear shift in the expectations for those designing

studies from including one power analysis for the main effect of treatment to also including power analyses for important moderator effects.

This shift toward including power analyses for moderator effects in efficacy trials represents unchartered territory for many researchers planning CRTs to test the efficacy of educational interventions. From 2002 to 2017, the focus was on designing CRTs to detect the main effect of treatment of a reasonable magnitude. During this time, several design principles emerged related to power for the main effect of treatment in CRTs. For instance, from a sample size perspective, it is now well known that the total number of schools is the key driver for increasing the power to detect the main effect of treatment in a CRT (e.g., Hedges & Rhoads, 2009; Raudenbush, 1997; Schochet, 2008). We also know that the smaller the intraclass correlation coefficient (ICC), or percentage of variance in the outcome that is between schools, the greater the power to detect main effects. Likewise, the inclusion of covariates that are strongly related to the outcome can increase the power to detect main effects. The empirical literature also suggests that educational interventions designed to improve achievement often yield effect sizes of approximately .20 to .30 standard deviation units, and hence powering a study to detect an effect of this magnitude is important (Hill et al., 2008). However, these same types of design principles and empirical analyses do not exist for power analyses for moderator effects in CRTs. But, given the changes in the RFA, it is important to start to investigate the design principles and the power of CRTs to detect moderator effects.

In this article, we examine the power to detect main effects and moderator effects for a sample of CRTs funded by IES. We intentionally select CRTs funded after 2012 when the RFA was modified to include more attention on moderator analyses. These studies represent CRTs that are typical in size and fall within budgetary constraints. We begin by asking the question:

1. What is the minimum detectable effect size (MDES) or the smallest effect size each study is designed to detect with a power of .80 (addressing the *what works* question)?

Given the emphasis on designing studies to detect treatment effects in the range of .20 to .30, and our use of planned sample sizes, we anticipate this will be the range for the MDES. Then we ask the following:

2. What is the minimum detectable effect size difference (MDESD) or the smallest differential effect size each study is designed to detect with a power for .80 for
   (a) Individual-level moderator effects ($MDESD_{IND}$; typically addressing the *for whom* question);
   (b) Cluster-level moderator effects ($MDESD_{CL}$; typically addressing the *under what conditions* question)?

Currently, no set of empirical benchmarks exist for the magnitude of moderator effects like the empirical benchmarks we rely on for main effects. Hence, we begin by simply determining the magnitude of the moderator effects studies are powered to detect and comparing it with the magnitude of the main effect studies are powered to detect. By considering the MDESDs and the MDES for the same set of studies, we are also able to examine where the design principles underlying power for main effects and moderator effects are consistent and where they diverge.

Our focus is specifically on design principles related to sample sizes. We focus on sample sizes for several reasons. First, the sample size is often something that is more under the control of the researcher than other design parameters. Second, our dataset includes studies with varying sample sizes at all levels which allows us to empirically examine the role of sample sizes. Third, we do not vary other design parameters across studies—that is, we use the same range for the ICC(s) and percentage of variance in the outcome that is explained by covariate(s). We use estimates for these design parameters based on recent empirical work, a practice which is common among researchers planning CRTs.

It is important to keep in mind that the CRTs in this sample were not required to be powered to detect moderator effects of a particular magnitude. Rather, these studies are being used to represent the typical size of efficacy trials funded by IES. The findings from this study will help the

field assess the potential for current CRTs and future CRTs that are similar in size and scope to answer questions about *what works*, *for whom*, and *under what conditions*. The findings will also help inform dialogues between funders and researchers about the feasibility of designing CRTs to sufficiently address all three types of questions. Note that this is a nontechnical presentation of statistical power and design implications and we refer readers to relevant literature throughout for a detailed discussion of the technical details behind the power calculations.

The article is organized as follows. We begin with a description of the sample of studies we used for the empirical analyses. Then we provide a brief overview of how we calculated the MDES, $MDESD_{IND}$ for an individual-level moderator, and $MDESD_{CL}$ for a cluster-level moderator. We present the findings for the studies and elucidate the role of sample sizes at different levels in power for main effects and moderator effects. We also consider the magnitude of the main effects and moderator effects the studies are designed to detect in light of what is known about effect sizes in education. Finally, we summarize the findings and consider the next steps in the quest to answer *what works*, *for whom*, and *under what conditions*.

Sample Description

The sample for this study included IES-funded efficacy trials between 2013 and 2018. We intentionally selected a starting point after 2012, 1 year after moderators started to be emphasized in the RFA. We identified the studies through the IES website (https://ies.ed.gov/funding/grantsearch/index.asp). There are four centers within IES: the National Center for Education Research (NCER), the National Center for Special Education Research (NCSER), the National Center for Education Evaluation and Regional Assistance (NCEE), and the National Center for Education Statistics (NCES). NCER funds the largest number of efficacy studies of the four IES centers and hence is the focus of this study. We searched funded grants to identify all of the efficacy studies funded by NCER between 2013 and 2018, a total of 75 studies (https://ies.ed.gov/funding/grantsearch/). For each efficacy study, we obtained the structured abstract. The structured abstract includes key information about each study related to Research Design and Methods, Key Measures, Data Analytic Strategy, Setting, and so on. The grantee is responsible for completing the structured abstracts at the time a study is funded, and hence the information in the structured abstract is based on the planned study. We selected a subsample to narrow down the sample to studies that were comparable.

Our inclusion criteria included the following: First, the study targeted students in grades pre-K–12 and academic achievement was one of the primary outcome variables. This removed two postsecondary studies and two additional studies that did not focus on academic outcomes. The design parameters, for example, ICC(s) and percentage of variance in the outcome that is explained by covariate(s), for planning studies focused on improving academic outcomes for grades pre-K–12 are often quite different from those focused on planning postsecondary studies or pre-K–12 studies focused on improving social–emotional or other types of outcomes (Bloom et al., 2007; Dong, Reinke, et al., 2016; Hedges & Hedberg, 2007, 2013; Westine et al., 2013). Hence, as we wanted to hold the range of design parameters constant across studies to isolate the effect of the varying sample sizes, it made sense to remove these studies. Furthermore, effect sizes are context specific and the magnitude of an effect in an academic domain in pre-K–12 may have a different meaning than that in a social–emotional or other domain or at the postsecondary level (Hill et al., 2008). Thus, for consistency and to enable comparisons across studies, we focused only on studies for pre-K–12 with academic achievement as a primary outcome.

Second, we restricted the sample to nested two- and three-level CRTs only. That is, we did not include multisite CRTs, for example, designs in which students are nested within schools and schools are randomly assigned to condition within multiple districts. To date, the literature and tools for calculating power for moderator effects in CRTs are limited. This study draws heavily on two papers that provide power calculations for moderator effects in two- and three-level CRTs (Dong et al., 2018; Spybrook et al., 2016). There is also a tool, PowerUp!-Moderator (Dong, Kelcey, et al., 2016), which allows users to conduct power calculations for moderator

effects for two- and three-level CRTs. However, the literature and tools for multisite CRTs are not as developed. Bloom and Spybrook (2017) examine power for moderator effects in multisite CRTs. However, they only consider site-level moderators. That is, they do not consider individual-level moderators or cluster-level moderators. Furthermore, PowerUp!-Moderator (Dong, Kelcey, et al., 2016) does not yet include options for calculating moderator effects for multisite CRTs. Given that we are demonstrating power calculations, we wanted to focus on designs with the literature and tools currently available for users planning CRTs and wanting to employ power calculations for moderator effects. As such, we removed 28 multisite CRTs.

Third, the study had to be an original study and not a follow-up of a prior CRT. Follow-up studies are intended to assess the longer-term outcomes of an intervention and often do not tend to include a CRT. Five follow-up studies were removed. Finally, one study was removed because sample sizes were not available via the structured abstract. Of the 75 studies originally identified as Cohort 2, 37 are included in this study. See Appendix for a list of the studies in the sample.

The topics, grade levels, and design classifications (discussed in the next section) for the 37 studies are identified in Table 1. From Table 1, we can see that Social, Behavioral, and Developmental interventions were the most common in this sample. The remaining topic areas were very similar in terms of the number of studies in the sample. Table 1 also revealed that the majority of the studies targeted students in elementary schools followed by pre-K.

Design Classification

The MDES and MDESD calculations differ slightly depending on whether the study is a two- or a three-level CRT. Hence, it is critical to first classify the study design. In a two- or three-level CRT, random assignment occurs at the top level. The difference in these two designs occurs in the total number of levels, 2 or 3. In more concrete terms, a two-level CRT may include students nested within schools in which schools are the unit of random assignment. Students represent Level 1 and schools represent Level 2, the top level and unit of randomization. A three-level CRT may include students nested within teachers nested within schools in which schools are again the unit of random assignment. Students represent Level 1, teachers represent Level 2, and schools represent Level 3, the top level and unit of randomization.

We classified the designs for the 37 studies in this sample. In addition, we identified the specific levels in each study. The design classifications are shown in Table 1. Approximately 73% ($n = 27$) of the studies were two-level CRTs. Of the two-level CRTs, the majority were designs in which students were nested within teachers or classrooms ($n = 17$). For the purposes of this article, we use the term teacher rather than classroom. If a teacher had more than one section, we refer to those as class sections. For these 27 studies, schools were not explicitly mentioned in the structured abstracts. Hence, it may be the case that multiple teachers were within one school or that there was only one teacher within each school. Without further information about the number of schools and distribution of teachers within schools, it was safest to assume one teacher per school. This provides more conservative estimates of the MDES and MDESD as the power to detect effects will often increase if there is some type of blocking of teachers within schools. Ten two-level CRTs included a nesting structure within students nested within schools. In these cases, the teacher level was not explicitly mentioned in the study abstract. Approximately 27% ($n = 10$) of the studies were three-level CRTs. These 10 studies followed the nesting structure of students nested within teachers nested within schools.

Sample Sizes

For each study, we determined the sample sizes at each level based on the structured abstract. The structured abstracts are written by the grantee after the study is funded, and hence the sample sizes reflect planned sample sizes. This is aligned with the purpose of this study which is to examine the MDES and MDESDs at the planning phase. In the few cases in which more than one treatment was randomly assigned, we calculated the total number of clusters for a two-group comparison. For example, consider a study with 80 schools and four conditions,

TABLE 1

Topic, Grade Level, and Design Classification for Studies in the Sample

	No. of studies (percent of total)
Topic	
Math and Science	7 (19)
Social, Behavioral, and Developmental	13 (35)
Literacy, Reading, and Writing	4 (11)
Teacher Quality and Professional Development	7 (19)
Other[a]	6 (16)
Grade level	
Pre-K	11 (30)
Elementary	18 (49)
Middle school	6 (16)
High school	2 (5)
Two-level CRT	
Students nested within teachers	17 (46)
Students nested within schools	10 (27)
Three-level CRT	
Students nested within teachers nested within schools	10 (27)

Note. CRT = cluster randomized trial.
[a]Other topics include English Learners, Educational Technology, Early Learning Programs, and State and Local Evaluations.

including three treatment conditions and one comparison condition. Assuming there were 20 schools per condition, we conducted the MDES and MDESD calculations using a total of 40 schools, 20 per each of the two conditions. In the cases where recruitment occurred across multiple years, we used the total sample size across years for the calculations.

The sample sizes for each of the 37 studies are provided in Table 2. In addition, the target grade for each study is also identified. The average number of schools in the three-level CRTs was 58 (median = 63). Similarly, the average number of schools in the two-level CRTs which randomized at the school level was 57 (median = 52). The average number of teachers in the two-level CRTs which randomized at the teacher level was 130 (median = 105). A closer look at Table 2 reveals that the majority of the two-level CRTs which randomized at the teacher level were pre-K studies. Pre-K classrooms are not necessarily housed within larger schools which may be partly why there were so many pre-K studies which randomized at the teacher level. It is also interesting to note that the studies with a small number of students either per teacher or per school, for example, 10 or fewer students per teacher or per school, tended to be pre-K studies. Those that were not pre-K studies but still had a small number of students per teacher or per school either served a special population or had special individualized testing circumstances that likely required one-on-one testing which is resource intensive.

Method

Next, we describe the MDES and MDESD calculations for a two- and a three-level CRT. Table 3 provides all of the formulas. More details and derivations for the formulas can be found in Bloom (2005), Spybrook et al. (2016), and Dong et al. (2018).

MDES (Addressing the What Works *Question)*

We begin with the MDES for the two-level CRT. As we can see from Table 3, to calculate the MDES, we need to know the total number of clusters, J, the approximate number of individuals per cluster, n, and the proportion of clusters randomly assigned to condition, P. For each study, we determined the sample sizes from the information obtained in the structured abstract. Across all studies, we assumed equal allocation of clusters to condition, or a 50–50 split. This assumption represents the ideal case and yields the smallest MDES. As a design moves away from the balanced case, the precision will decrease. However, it is important to be aware that small deviations, such as a 60–40 split, will not result in major changes to the MDES.

From Table 3, we can also see that the MDES depends on an estimate of the proportion of variance between clusters, ICC, and an estimate of the proportion of variance explained by covariates, R^2_{L1} and R^2_{L2}. These values are not included in the online structured abstracts. However, in the past decade, we have seen an emerging empirical database of design parameters necessary for

TABLE 2

Sample Sizes and Grade Level for Each Study in the Sample

Three-level CRT (Study ID)	Total no. of schools	Avg. no. of teachers per school	Avg. no. of students per teacher	Target grade
1	72	3	10	Pre-K
2	30	20	23	Middle
3	66	4	52	Middle
4	32	8	25	Elementary
5	30	16	65	Elementary
6	56	3	25	Elementary
7[a]	60	7	6	Elementary
8[a]	66	6	20	Elementary
9[a]	70	5	3	Elementary
10[a]	100	2	4	Elementary
Two-Level CRT (Study ID)		**Total no. of schools**	**No. of students per school**	**Target grade**
11		20	24	Elementary
12		30	21	Elementary
13		52	673	Elementary
14		56	20	Elementary
15		81	60	Elementary
16		85	200	Elementary
17[a]		40	18	Middle
18[a]		50	10	Middle
19[a]		52	10	High
20		103	100	High
Two-Level CRT (Study ID)		**Total no. of teachers**	**No. of students per teacher**	**Target grade**
21		60	8	Pre-K
22		60	8	Pre-K
23		64	19	Pre-K
24		100	8	Pre-K
25		100	10	Pre-K
26		120	10	Pre-K
27		120	10	Pre-K
28		120	18	Pre-K
29		140	8	Pre-K
30		220	5	Pre-K
31		84	26	Elementary
32[a]		110	25	Elementary
33		160	19	Elementary
34		440	10	Elementary
35		130	40	Elementary
36		55	20	Middle
37		100	25	Middle

Note. CRT = cluster randomized trial.

[a]Studies represent those included a special population, for example, English Language Learners, students with severe social anxiety, social skill challenges, disruptive behaviors, or special individualized testing circumstances.

TABLE 3

Detailed Formulas for MDES, $MDESD_{IND}$, and $MDESD_{CL}$ for Two- and Three-Level CRTs

	Two-level CRT	Three-level CRT
MDES	$\mathrm{MDES}_{2\mathrm{LCRT}} = \frac{M_{J-3}}{\sqrt{J}}\sqrt{\left(1-R^2_{L2}\right)\rho + \frac{\left(1-R^2_{L1}\right)\left(1-\rho\right)}{n}}\sqrt{\frac{1}{P\left(1-P\right)}}$, where n is the number of individuals per cluster; J is the total number of clusters; and M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $J - 3$ degrees of freedom. If the degree of freedom is greater than 20, M is approximately 2.8 (for more details, please see Bloom, 1995); ρ is the intraclass correlation (ICC) that is defined as $\tau_{L2}/(\tau_{L2}+\tau_{L1})$ and τ_{L2} is the variance between clusters and τ_{L1} is the variance within clusters; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; R^2_{L2} is the proportion of variance explained by a single Level 2 covariate; and P is the proportion of clusters assigned to treatment that we assume to be .50 (Bloom, 2005; Spybrook et al., 2016).	$\mathrm{MDES}_{3\mathrm{LCRT}} = \frac{M_{K-3}}{\sqrt{K}}\sqrt{\left(1-R^2_{L3}\right)\rho_{L3} + \frac{\left(1-R^2_{L2}\right)\rho_{L2}}{J} + \frac{\left(1-R^2_{L1}\right)\left(1-\rho_{L3}-\rho_{L2}\right)}{Jn}}\sqrt{\frac{1}{P\left(1-P\right)}}$, where n is the number of individuals per cluster; J is the number of Level 2 subclusters per Level 3 cluster; K is the total number of Level 3 clusters; M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $J - 3$ degrees of freedom; ρ_{L3} is the ICC at Level 3 that is defined as $\tau_{L2}/(\tau_{L3}+\tau_{L2}+\tau_{L1})$ and τ_{L3} is the variance between Level 3 clusters, τ_{L2} is the variance between Level 2 subclusters, and τ_{L1} is the variance within subclusters; ρ_{L1} is the ICC at Level 2 that is defined as $\tau_{L2}/(\tau_{L3}+\tau_{L2}+\tau_{L1})$ and τ_{L3} is the variance between Level 3 clusters, τ_{L2} is the variance between Level 2 subclusters, and τ_{L1} is the variance within subclusters; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; R^2_{L2} is the proportion of variance explained by a single Level 2 covariate; R^2_{L3} is the proportion of variance explained by a single Level 3 covariate; and P is the proportion of clusters assigned to treatment that we assume to be .50 (Dong et al., 2018).
$\mathrm{MDESD}_{\mathrm{IND}}$	$\mathrm{MDESD}_{\mathrm{IND(2LCRT)}} = \frac{M_{Jn-J-2}}{\sqrt{J}}\sqrt{\frac{\left(1-R^2_{L1}\right)\left(1-\rho\right)}{n}}\sqrt{\frac{1}{P\left(1-P\right)Q\left(1-Q\right)}}$ where n is the number of individuals per cluster; J is the total number of clusters; and M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $Jn - J - 2$ degrees of freedom; ρ is the ICC; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; P is the proportion of clusters assigned to treatment that we assume to be .50; and Q is the proportion of individuals in each moderator subgroup which we assume to be .50 (Bloom, 2005; Spybrook et al., 2016).	$\mathrm{MDESD}_{\mathrm{IND(3LCRT)}} = \frac{M_{KJn-KJ-K-2}}{\sqrt{K}}\sqrt{\frac{\left(1-R^2_{L1}\right)\left(1-\rho_{L3}-\rho_{L2}\right)}{Jn}}\sqrt{\frac{1}{P\left(1-P\right)Q\left(1-Q\right)}}$ where n is the number of individuals per cluster; J is the number of Level 2 subclusters per Level 3 cluster; K is the total number of Level 3 clusters; M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $KJn - KJ - K - 2$ degrees of freedom; ρ_{L3} is the ICC at Level 3; ρ_{L2} is the ICC at Level 2; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; P is the proportion of clusters assigned to treatment that we assume to be .50; and Q is the proportion of individuals in each moderator subgroup which we assume to be .50 (Dong et al., 2018).

(continued)

TABLE 3 (continued)

	Two-level CRT	Three-level CRT
$MDESD_{CL}$		$MDESD_{TCHR(3LCRT)} = \frac{M_{KJ-K-2}}{\sqrt{K}}\sqrt{\frac{\left(1-R^2_{L2}\right)\rho_{L2}}{J}+\frac{\left(1-R^2_{L1}\right)\left(1-\rho_{L3}-\rho_{L2}\right)}{Jn}}\sqrt{\frac{1}{P(1-P)Q(1-Q))}}$ where n is the number of individuals per cluster; J is the number of Level 2 subclusters per Level 3 cluster; K is the total number of Level 3 clusters; M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $KJ - K - 2$ degrees of freedom; ρ_{L2} is the ICC at Level 2; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; R^2_{L2} is the proportion of variance explained by a single Level 2 covariate; P is the proportion of clusters assigned to treatment that we assume to be .50; and Q is the proportion of subclusters in each moderator subgroup that we assume to be .50 (Dong et al., 2018).
$MDESD_{CL}$	$MDESD_{CL(2LCRT)} = \frac{M_{J-5}}{\sqrt{J}}\sqrt{\left(1-R^2_{L2}\right)\rho+\frac{\left(1-R^2_{L1}\right)(1-\rho)}{n}}\sqrt{\frac{1}{P(1-P)Q(1-Q)}}$ where n is the number of individuals per cluster; J is the total number of clusters; and M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $J - 3$ degrees of freedom; ρ is the ICC; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; R^2_{L1} is the proportion of variance explained by a single Level 2 covariate; P is the proportion of clusters assigned to treatment that we assume to be .50; and Q is the proportion of clusters in each moderator subgroup that we assume to be .50 (Dong et al., 2018).	$MDESD_{SCHL(3LCRT)} = \frac{M_{K-5}}{\sqrt{K}}\sqrt{\left(1-R^2_{L3}\right)\rho_{L3}+\frac{\left(1-R^2_{L2}\right)\rho_{L2}}{J}+\frac{\left(1-R^2_{L1}\right)\left(1-\rho_{L3}-\rho_{L2}\right)}{Jn}}\sqrt{\frac{1}{P(1-P)Q(1-Q))}}$ where n is the number of individuals per cluster; J is the number of Level 2 subclusters per Level 3 cluster; K is the total number of Level 3 clusters; M is the group effect multiplier that corresponds to the value of the t distribution for a two-tailed test with $\alpha = .05$, power = .80, equal variances for groups, and $K - 5$ degrees of freedom; ρ_{L3} is the ICC at Level 3; ρ_{L2} is the ICC at Level 2; R^2_{L1} is the proportion of variance explained by a single Level 1 covariate; R^2_{L1} is the proportion of variance explained by a single Level 2 covariate; R^2_{L3} is the proportion of variance explained by a single Level 3 covariate; P is the proportion of clusters assigned to treatment that we assume to be .50; and Q is the proportion of clusters in each moderator subgroup that we assume to be .50 (Dong et al., 2018).

Note. CRT = cluster randomized trial; MDES = minimum detectable effect size; MDESD = minimum detectable effect size difference.

TABLE 4

Design Parameters Used in Calculating MDES, $MDESD_{IND}$, and $MDESD_{CL}$ for Two- and Three-Level CRTs

	Two-level CRT			Three-level CRT				
	ICC	R^2_{L1}	R^2_{L2}	ICC_3	ICC_2	R^2_{L1}	R^2_{L2}	R^2_{L3}
MDES	.15, .25	.20, .50	.50, .80	.10, .15	.07, .10	.20, .50	.20, .50	.50, .80
$MDESD_{IND}$	.15, .25	.20, .50	—	.10, .15	.07, .10	.20, .50	—	—
$MDESD_{CL*}$	—	—	—	.10, .15	.07, .10	.20, .50	.20, .50	—
$MDESD_{CL}$	.15, .25	.20, .50	.50, .80	.10, .15	.07, .10	.20, .50	.20, .50	.50, .80

Note. $MDESD_{CL}$ in Row 3 for the three-level CRT corresponds to the teacher-level moderator and $MDES_{CL}$ in Row 4 for the three-level CRT corresponds to the school-level moderator. Estimates are based on Bloom et al. (2007), Brandon et al. (2013), Hedges and Hedberg (2007, 2013), Jacob et al. (2010), Online Variance Almanac (n.d.), Spybrook et al. (2016), Westine et al. (2013), and Zhu et al. (2012). MDES = minimum detectable effect size; CRT = cluster randomized trials; ICC = intraclass correlation.

planning CRTs, particularly for CRTs focused on academic achievement. This set of empirical estimates is often used for planning CRTs and, as such, we used the empirical literature to estimate the ICC, R^2_{L1}, and R^2_{L2}.

The empirical literature suggests that design parameters vary by context, where context includes factors such as grade level, subject area, and types of schools. To account for this variation, we use a range of values to estimate the design parameters. Note that the upper and lower bounds we use for the empirical estimates of the design parameters capture the typical variations across grade level, subject areas, types of schools, and so on when the outcome is academic achievement. We estimate the design parameters for the two-level CRTs from the empirical studies which nest students within schools.

Although 17 of our studies actually have teacher as Level 2, recall that we assumed one teacher per school. As such, school and teacher are confounded and estimates of design parameters from data with students nested within schools are reasonable. The estimates we used for the ICC, R^2_{L1}, and R^2_{L2} are based on the range of estimates found in the empirical literature and are provided in Table 4 (e.g., Bloom et al., 2007; Brandon et al., 2013; Hedges & Hedberg, 2007, 2013; Jacob et al., 2010; Spybrook et al., 2016; Westine et al., 2013; Zhu et al., 2012). In our calculations, we assume one covariate at each level. We could include multiple covariates at each level. However, each additional covariate at Level 2 results in the loss of one additional degree of freedom. Given that the pretest is a common and powerful covariate in CRTs focused on academic achievement and that, after the pretest is included, additional covariates do not tend to explain much more variation, we assume that the covariate is a pretest and do not include other covariates.

Looking to the right in Table 3, we see the MDES for a three-level CRT. The MDES looks very similar to the two-level CRT. The key difference is that now there are three sample sizes, two ICCs, and potentially three R^2 values. As in the case of the two-level CRT, we used the structured abstract to determine the relevant sample sizes for each study and assumed a balanced design. We turned to the empirical literature to estimate the ICCs and R^2 values (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007, 2013; Jacob et al., 2010; Westine et al., 2013). Although there is quite a substantial literature base of empirical estimates of ICCs for two-level studies with students nested within schools, there is much less available for three-level studies with students nested within teachers nested within schools.

To our knowledge, there are three studies that estimate design parameters for students nested within teachers nested within schools (Jacob et al., 2010; Nye et al., 2004; Xu & Nichols, 2010). For reading and math outcomes, these studies tended to report approximately 5% to 10% of the variation in the outcome at the teacher level and large portions of teacher- and school-level variance explained by pretests. We use this to guide our estimates of the ICC ranges as shown in Table 4.

$MDESD_{IND}$ *(Addressing the* for Whom *Question)*

Similar to the MDES, the $MDESD_{IND}$ calculations differ depending on the design. We begin with the two-level CRT. For illustrative purposes, we assume a binary individual-level moderator, such as gender. From Table 3, we can see several differences in the $MDESD_{IND}$ formula compared with the MDES formula. For example, the Level 2 variance and the percent of variance explained at Level 2 do not factor into the $MDESD_{IND}$ calculations. This is because in the case of an individual-level moderator, such as gender, in a two-level design, the differences in boys and girls are within clusters and thus the school effects cancel out (Spybrook et al., 2016). This is critical because, as discussed earlier, research over the past 15 years has established that the ICC plays a big role in the MDES calculations. That is, the larger the ICC, the larger the MDES (e.g., Hedges & Rhoads, 2009; Raudenbush, 1997; Schochet, 2008). Furthermore, as the school effects cancel out, the total sample size, $n \times J$, becomes the critical sample size, whereas the MDES is largely driven by the total number of clusters. Another important difference is that, in addition to P, the proportion of clusters assigned to condition, we need to also specify Q, the proportion of individuals in each moderator subgroup. Throughout our calculations, we assume an equal proportion of individuals in each subgroup which again represents the ideal case. For moderators such as gender, this may be a realistic assumption. For other moderators, such as free or reduced price lunch status, this may not be appropriate and we caution researchers to consider this carefully. Similar to the allocation of clusters to condition, the more imbalanced the design, the larger the $MDESD_{IND}$. The empirical estimates of the relevant design parameters for $MDESD_{IND}$ are shown in Table 4. Although inclusion of the moderator may explain some additional variance at the level of the moderator, we use the same range of estimates of R^2_{L1} as a conservative lower bound. We do not estimate R^2_{L2} as this does not enter the calculations and the ICC is necessary only because the Level 1 variance contributes to the calculations of $MDESD_{IND}$.

$MDESD_{IND}$ for the three-level CRT follows the same pattern as we saw in the case of the two-level CRT. That is, the between-school and between-teacher effects cancel out so the variance components at these levels do not contribute to the variance of the moderator effect and the total sample size becomes the key driver of $MDESD_{IND}$. Again, we assume an equal proportion of schools assigned to each condition and an equal proportion of individuals in each subgroup. As shown in Table 4, the relevant design parameters for $MDESD_{IND}$ for the three-level CRT are consistent with those used to estimate the MDES.

$MDESD_{CL}$ *(Addressing the* Under What Conditions *Question)*

As in the case of an individual-level moderator, we assume a binary cluster-level moderator. In the two-level CRTs, we have both schools and teachers as clusters. As such, the moderator might be something like school type (urban or rural) if the cluster is school or teacher experience (new vs. veteran) if the cluster is teacher. As the cluster may represent schools or teachers, we simply refer to it as the cluster for the two-level CRT knowing that it may represent a school- or a teacher-level moderator.

In Table 3, we provide the formula for $MDESD_{CL}$ for the two-level CRT in the second row of the $MDESD_{CL}$ values for consistency with the three-level CRT because the moderator in this case is at the top level. The equation for $MDESD_{CL}$ looks very similar to that of the MDES for a two-level CRT. As such, similar to MDES, the total number of clusters will be the key driver of $MDESD_{CL}$. There is also the addition of the Q term. Assuming an equal proportion of clusters assigned to condition and an equal proportion of clusters in each cluster-level moderator subgroup, this suggests that, holding all else constant, $MDESD_{CL}$ will be larger than the MDES. Table 4 shows the same empirical estimates used for $MDESD_{CL}$. As in the case of $MDESD_{IND}$, we use the same range of estimates for the R^2 values and the ICC noting that the inclusion of the moderator may explain some additional variance at Level 2 and as such we are estimating a conservative lower bound.

In a three-level CRT, there are two levels of clustering. In our studies, teachers are at Level 2, whereas schools are at Level 3. We calculate $MDESD_{CL}$ for moderators at both levels and denote them

$MDESD_{TCHR(3LCRT)}$ and $MDESD_{SCHL(3LCRT)}$, respectively. We begin by looking at $MDESD_{SCHL(3LCRT)}$ in Table 3. This equation looks very similar to that of the MDES for a three-level CRT with the addition of the Q term. As such, it will function similar to the MDES where the total number of schools will be the key sample size. However, $MDESD_{TCHR(3LCRT)}$ looks slightly different than the MDES and $MDESD_{SCHL(3LCRT)}$. Although it also has a Q term, Q in this case represents the proportion of teachers in the moderator subgroup. Furthermore, the between-school variance cancels out, similar to the case of the individual-level moderator. As such, the between-school variance does not contribute to the calculations and the total number of teachers becomes the critical sample size. Table 4 shows the same empirical estimates used for $MDESD_{TCHR(3LCRT)}$ and $MDESD_{SCHL(3LCRT)}$. As in the case of $MDESD_{IND}$, we use the same range of estimates for the relevant R^2 values and the ICC noting that the inclusion of the moderator may explain some additional variance and as such we are estimating a slightly conservative lower bound.

Results

We begin with the results for the MDES. Then we present the findings for $MDESD_{IND}$ for the individual-level moderators. Next, we consider cluster-level moderators. We present the findings for the MDESD for a school-level moderator followed by the findings for a teacher-level moderator. The MDES and MDESD are graphed together to facilitate comparisons. Furthermore, Study ID in all of the figures matches Study ID in Table 2.

MDES (Addressing the What Works *Question)*

The MDES for each of the 37 studies is shown by the striped bars in Figure 1. A range of the MDES is graphed for each study because a range of design parameters was used for all the calculations to account for the variability in design parameters. The mean of the midpoint of the MDES across studies is .21 (SD = .06). As we expected, this finding is consistent with benchmarks for meaningful effect sizes in intervention studies focused on improving academic outcomes suggested by Hill et al. (2008).

Hill et al. (2008) examined 61 randomized studies and found average effect sizes ranging from .27 to .51 for interventions designed to improve achievement outcomes from elementary through high school grades. Furthermore, they examined 76 meta-analyses of educational interventions and found average effect sizes ranging between .20 and .30. In general, they suggested that studies should be designed to detect effect sizes for the mean effect of treatment in the range of .20 to .30. It is interesting to note that the IES RFA does not specify the MDES for a study. Rather, the RFA specifies that one should conduct a power analysis and provide a strong rationale for the appropriateness of the magnitude of the main effect the study is powered to detect. The findings in Figure 1 suggest that most studies are designed with power to detect main effects in a reasonable range based on empirical benchmarks.

$MDESD_{IND}$ (Addressing the for Whom *Question)*

Now that we know that most studies are designed to detect a main effect of a reasonable magnitude, the next question is what is the magnitude of individual moderator effects that these same studies are powered to detect. Regardless of whether a study is a two- or a three-level CRT, $MDESD_{IND}$ can be calculated. The results of $MDESD_{IND}$ for all 37 studies are displayed by the solid bars in Figure 1.

In general, the range for $MDESD_{IND}$ is smaller than that of the MDES because as discussed earlier the school effects cancel out. As such, the power calculations are simplified. For example, the variance explained at Level 2, R^2_{L2}, which introduces additional variance into the MDES calculations does not factor into the $MDESD_{IND}$ calculations. The mean for $MDESD_{IND}$ is .19 (SD = .10). Recall the mean of the MDES is .21. Although the means are similar, from Figure 1 it is clear that there are some cases in which $MDESD_{IND}$ is smaller than the MDES and others where it is larger than the MDES. So the question is what is driving these differences.

Let us consider Studies 13 and 21. From Figure 1, we can see that the range of the MDES for Studies 13 and 21 was approximately .14 to .28 and .21 to .33, respectively. The range of

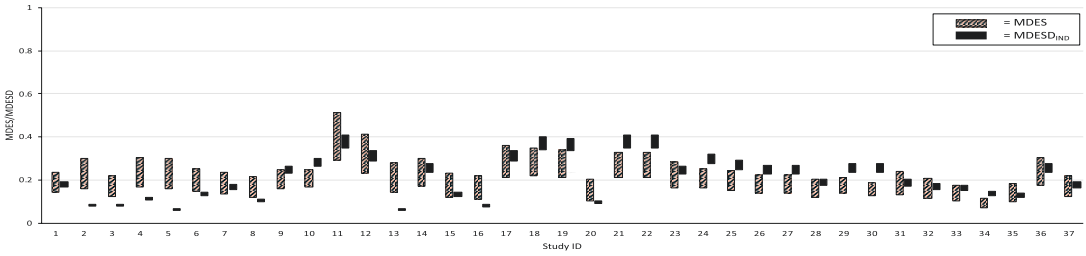

FIGURE 1. *Minimum detectable effect size (MDES) and minimum detectable effect size difference (MDESD) for individual-level moderators for all studies.*

Note. The MDES is represented by striped bars and appears to the left of Study ID. The MDESD is represented by solid bars and appears to the right of Study ID. The MDES and MDESD are ranges based on the assumptions in Table 2 for the sample sizes and Table 4 for the design parameters.

$MDESD_{IND}$ for Studies 13 and 21 was approximately .03 to .05 and .33 to .39, respectively. So although the range for the MDES for the two studies was not that different, the range for $MDESD_{IND}$ was very different. As they both were based on the same estimates of the design parameters, we know that the differences are a function of the sample sizes. From Table 2, we can obtain the sample sizes for both studies. Study 13 was a two-level CRT with approximately 673 students per school and a total of 52 schools. Study 21 was a two-level CRT with approximately eight students per teacher and a total of 60 teachers. The two studies have a similar number of total clusters, 52 and 60, respectively. As the MDES is driven by the total number of clusters, it makes sense that the range of the MDES is similar for the two studies. However, they are very different in terms of the number of students per cluster, 673 and 8, respectively. $MDESD_{IND}$ is driven by the total number of students and thus Study 13, with a total of $673 \times 52 = 34{,}996$ students relative to Study 21, with a total of $8 \times 60 = 480$ students, has the capacity to detect much smaller individual-level moderator effects.

From Table 2, we can see that the studies with smaller numbers of students per cluster tend to be the pre-K studies, Studies 21 to 30, and in some cases studies that randomized teachers, Studies 31 to 37. Studies with large numbers of total students tend to be elementary, middle, and high school studies that randomize schools, Studies 1 to 20. These include two-level CRTs with students nested within schools, Studies 11 to 20, or three-level CRTs with students nested within teachers nested within schools, Studies 1 to 10. Note that for a three-level CRT the total number of students is the number of students per teacher multiplied by the number of teachers per school multiplied by the total number of schools. Figure 1 reflects the importance of large numbers of individuals per cluster in decreasing $MDESD_{IND}$ relative to the MDES as Studies 1 to 20, or studies which randomize schools, reveal cases where $MDESD_{IND}$ is smaller than the MDES.

It is also important to consider the magnitude of the moderator effects these studies are designed to detect. Given the lack of empirical benchmarks, we consider the magnitude of the moderator effect relative to the magnitude of the main effect. In the case of a binary moderator, the moderator effect represents a differential effect between two groups. Based on prior research from psychology, we anticipate that moderator effects will be smaller than the main effect (Aguinis et al., 2005). Hence, the studies which have a smaller $MDESD_{IND}$ than MDES, such as Study 13, will tend to be in a stronger position to detect individual-level moderator effects. As discussed above, this tends to be elementary, middle, and high school studies that randomize schools. Studies of pre-K interventions with small numbers of individuals per cluster are likely to not be able to detect reasonable individual-level moderator effects.

$MDESD_{CL}$ *(Addressing the* Under What Conditions *Question)*

We begin by examining the capacity of studies to detect school-level moderators. Hence, the studies that randomize at the school level, the two-level CRTs with schools at Level 2 and

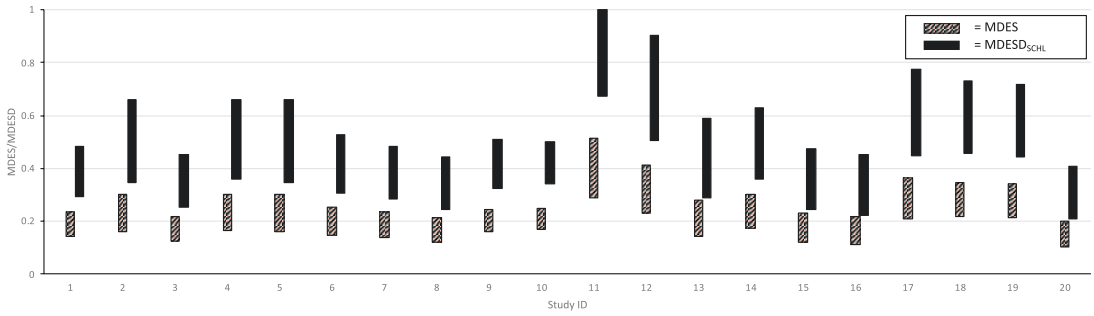

FIGURE 2. *Minimum detectable effect size (MDES) and minimum detectable effect size difference (MDESD) for a school-level moderator for the three-level CRTs and two-level CRTs with schools as the top level in the sample.*

Note. The MDES is represented by striped bars and appears to the left of Study ID. The MDESD is represented by solid bars and appears to the right of Study ID. The MDES and MDESD are ranges based on the assumptions in Table 2 for the sample sizes and Table 4 for the design parameters. The upper bound for Study 11 for the MDESD is truncated at 1.0.

the three-level CRTs in the sample are relevant. This includes a total of 20 studies. The findings are presented alongside the MDES for each study in Figure 2. The mean of $MDESD_{SCHL}$ is .48 (SD = .15).

Clearly, the magnitude of $MDESD_{SCHL}$ is quite a bit larger than that of the MDES. For example, consider Study 5, a study targeting elementary schools with a total of 30 schools, 16 teachers per school, and 65 students per teacher. The MDES ranges from .15 to .30. However, $MDESD_{SCHL}$ ranges from .35 to .66. This is because just like the MDES, $MDESD_{SCHL}$ is driven by the total number of schools. In the case of a binary moderator with equal allocation of schools to condition and equal numbers of schools per moderator subgroup, it is similar to a study that compares four groups rather than two groups and hence the magnitude of $MDESD_{SCHL}$ is approximately twice that of the MDES (Spybrook et al., 2016). Furthermore, the total number of schools is the key sample size unlike the case of $MDESD_{IND}$ where the number of students per school was extremely helpful in detecting smaller individual-level moderator effects.

Just like in the case of an individual-level moderator, the cluster-level moderator represents a differential treatment effect. Again, the lack of benchmarks for the magnitude of moderator effects makes it challenging to interpret these findings. However, if we assume that moderator effects will be smaller than the main effects, this suggests that current studies are not well positioned to detect meaningful school-level moderator effects. Even if we assume that cluster-level moderator effects may be similar to main effects, the findings in Figure 2 reveal that studies are not powered to detect moderator effects of similar magnitudes.

Next, we consider the teacher-level moderator. We begin with two-level CRTs that have teachers at Level 2 (n = 17). In these studies, teachers represent the top level. Hence, the calculations are identical to those performed for a two-level CRT with schools at the top level. We separate them out here because substantively it is different to think about teacher-level moderators, for example, teacher experience, than school-level moderators, for example, school type.

The findings are shown in Figure 3. The mean $MDESD_{TCHR(2LCRT)}$ value is .37 (SD = .10). From Figure 3, we can see that $MDESD_{TCHR(2LCRT)}$ is larger than the MDES. Again, this is because, just like in the case of the two-level CRT with schools at the top level, the MDESD is driven by the total number of clusters, or teachers in this case. With regard to the magnitude of the teacher-level moderator effect when teachers are at the top level, the fact that it is quite a bit higher than the main effect as can be seen in Figure 3 suggests that these studies may not have the capacity to detect meaningful teacher-level moderator effects.

We also consider the capacity of studies to detect teacher-level moderator effects in

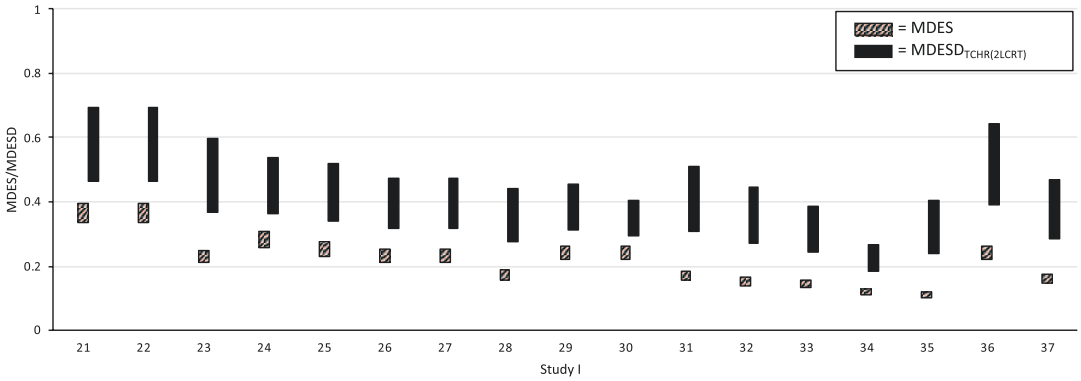

FIGURE 3. *Minimum detectable effect size (MDES) and minimum detectable effect size difference (MDESD) for a teacher-level moderator for the two-level CRTs with teachers as the top level in the sample.*
Note. The MDES is represented by striped bars and appears to the left of Study ID. The MDESD is represented by solid bars and appears to the right of Study ID. The MDES and MDESD are ranges based on the assumptions in Table 2 for the sample sizes and Table 4 for the design parameters.

three-level CRTs. In this case, the moderator is at the middle level, one level lower than the level of randomization. As shown in Table 3, the magnitude of $MDESD_{TCHR(3LCRT)}$ in the three-level CRT is driven by the total number of teachers, which is the number of teachers per school multiplied by the total number of schools.

The MDES and $MDESD_{TCHR(3LCRT)}$ values are shown in Figure 4. The mean $MDESD_{TCHR(3LCRT)}$ value is .21 (SD = .07). From Figure 4, we can see that in some cases $MDESD_{TCHR(3LCRT)}$ is smaller than the MDES. For example, in Study 2, there are 20 teachers per each of 30 schools. Hence, the calculations are based on a sample size of $20 \times 30 = 600$ teachers and $MDESD_{TCHR(3LCRT)}$ ranges from approximately .11 to .15.

In the case of a three-level CRT with teachers at Level 2, studies with large numbers of teachers per school may be able to detect smaller teacher-level moderator effects than main effects. From Table 2, we see one middle school study and one elementary study with 20 and 16 teachers, respectively. In terms of the magnitude of the teacher-level moderator effects the studies are able to detect, from Figure 4 we can see that several of the studies may be able to detect moderator effects that are smaller than .20, the typical magnitude we desire to detect for main effects. Again, without empirical benchmarks to guide these interpretations, it is hard to anticipate if these are reasonable; however, we do observe that they are smaller than what is deemed to be reasonable for main effects.

Conclusion

The findings from this article suggest that recently funded IES efficacy trials are designed to detect main effects of treatment of approximately .20 standard deviation units. Given the recent empirical literature which suggests that boosting academic achievement by .20 standard deviation units is a practically significant effect, we would expect studies to be designed to meet this threshold. As such, our results concur with other studies that IES-funded CRTs are well positioned to answer the *what works* question (Spybrook et al., 2016). The push toward also understanding *for whom* and *under what circumstances* an intervention is effective suggests the importance of assessing the capacity of studies to provide rigorous evidence of the effects of individual-, teacher-, and school-level moderators. The findings from this study shed light on these questions.

We begin with the *for whom* question. Overall, some studies were well positioned to detect student-level moderator effects that were less than .20, and in some cases as small as .03 to .05. Studies that were well positioned to detect smaller student-level moderator effects were those that randomized elementary, middle, or high schools and included larger numbers of

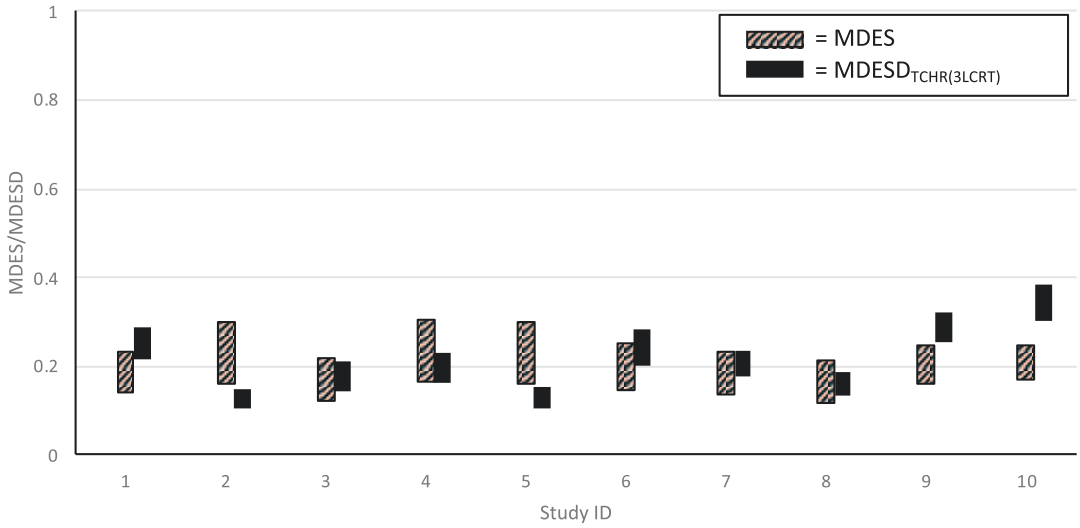

FIGURE 4. *Minimum detectable effect size (MDES) and minimum detectable effect size difference (MDESD) for a teacher-level moderator for the three-level CRTs in the sample.*
Note. The MDES is represented by striped bars and appears to the left of Study ID. The MDESD is represented by solid bars and appears to the right of Study ID. The MDES and MDESD are ranges based on the assumptions in Table 2 for the sample sizes and Table 4 for the design parameters.

students per school. This is a direct result of design principles for powering CRTs to detect student-level moderator effects. That is, although the number of clusters drives the magnitude of the main effects studies are designed to detect, the total number of individuals, number of clusters times the number of individuals per cluster, drives the magnitude of the student-level moderator effects studies are designed to detect. Hence, if a study is seeking to answer not only *what works* but also *for whom*, larger numbers of students per school will be helpful in answering the *for whom* question. This is important as increasing the number of students per school may not be very costly, particularly in elementary, middle, or high schools which often use standardized assessment as the key academic outcome.

However, some studies were only able to detect student-level moderators that were larger than .20, which is not likely to be seen in practice. These were pre-K studies or studies with special populations. Often, these studies are limited in the number of individuals per cluster. Or they rely on individual administration of tests, in which case it may be very costly to include more students.

The question of *under what circumstances* a treatment is effective, or questions related to school-level moderators, will be challenging to answer for current studies. The studies were all designed to detect school-level moderator effects larger than the main effects, which is not likely to be seen in practice. This is because, from a design perspective, the total number of clusters drives the power for both main effects and school-level moderator effects. Typically, there are enough schools in a study to be powered to detect a reasonable main effect, but recruiting additional schools to increase the capacity of a study to detect a school-level moderator can be very costly and likely outside of the budget for one study.

Questions related to teacher-level moderators are slightly more complicated as they depend partly on whether teachers are the unit of assignment. If they are, as in the case of a two-level CRT with students nested within teachers, the magnitude of teacher-level moderators these studies were designed to detect was approximately .40, which is quite large and not likely to be seen in practice. From the design perspective, much like school-level moderator effects in a school randomized study, the number of teachers drives the power for the main effect and the teacher-level moderator. However, in the case in which teachers represent a level below the unit of

assignment, for example, a three-level CRT with students nested within teachers nested within schools, studies with large numbers of teachers per school, some studies were able to detect moderator effects smaller than the main effects. From a design perspective, this is because teachers are at a level lower than the unit of randomization and hence the total number of teachers, the number of schools times the number of teachers, is the critical sample size. In practice, for large school-wide interventions where schools are randomized and there are many teachers per school, increasing the number of teachers may be a relatively inexpensive strategy to increase the capacity to detect teacher-level moderator effects. However, for studies that target particular grades within schools and hence are limited to the number of teachers in a specific grade, this may not be possible.

Limitations

There are several limitations to this study. First, it is important to keep in mind that these studies were designed to have adequate power to detect the main effect of treatment. This translates to many studies having power to detect an effect size of .20 standard deviation units. However, these studies were not required to be adequately powered to detect meaningful moderator effects. Hence, these studies are being used to represent typical IES studies to demonstrate likely levels of power to detect moderator effects.

Second, the sample sizes used in this study were obtained from online structured abstracts.[1] Structured abstracts are submitted after a study is funded. However, studies may change after the initial funding period and this may include changes to the design and sample sizes. Spybrook et al. (2013) examined changes in sample size and precision between the planning phase and the implementation phase of a set of early CRTs funded by IES and found that, in the majority of studies, changes in sample sizes between phases did not lead to major changes in the precision of the study. However, any changes that may have occurred in the sample size in this sample of studies are not accounted for in these analyses.

Third, some structured abstracts provided more information than others. In cases in which designs or sample sizes were not clear, we used our knowledge of research design to classify studies to the best of our ability. However, our assessments may not be consistent with the original intent of the authors.

Fourth, we used the empirical literature to estimate design parameters. Individual studies may have used different estimates of design parameters specific for their study and hence the use of the empirical literature may lead to an overestimate or underestimate of the MDES and MDESDs. Fifth, we assume equal allocation of teachers or schools to condition, and equal allocation of students, teacher, or schools to different moderator subgroups. Deviations away from this will lessen the power to detect main effects and moderator effects. Finally, we did not include multisite CRTs in our sample as the methodological work related to the power calculations for moderator effects in multisite CRTs lags behind that of two- and three-level CRTs. Given the limitations in this study, the findings are not meant to be definitive in nature. Rather, they are meant to help the field start to understand the likely capacity of typical size CRTs funded by IES to help answer questions about *what works*, *for whom*, and *under what conditions*.

Looking Forward

Moving beyond the *what works* question and considering questions about *for whom* and *under what conditions* an intervention is effective is critical in meeting the mission of IES to improve education outcomes for all students. These are important questions that will help policymakers and school personnel make better decisions about which interventions to adopt. We applaud IES for pushing researchers to answer these critical questions. However, we believe that it is important to also consider whether one study, given the typical resource allotment, can achieve all of these goals. Although we show that there may be potential for studies conducted in elementary, middle, and high schools to detect meaningful moderators at the student or teacher level, we also show that this is not likely in pre-K studies or studies of special populations where the number of students per teacher or school may be small. Furthermore, regardless of target grade, it is more challenging to design studies with the capacity to detect meaningful effects of school-level moderators.

Given some of the potential challenges associated with powering studies to detect effects of school-level moderators, perhaps IES might encourage greater collaboration across studies during the design phase of studies. For example, they might encourage or incentivize the collection of a set of common moderator variables and outcomes across studies. This would facilitate tests of moderator effects of similar outcomes across several studies where pooling across studies may lead to greater power to detect important moderator effects. On a larger scale, common moderator variables and outcomes would also lead to stronger meta-analyses of the effects of various interventions. Working together, particularly in the design phase of studies, will likely help us move closer to building a body of evidence on which to base education policy and practice, a central goal of IES.

More support for methodological work related to power for moderator effects is also critical. Over the past several years, we have started to see more work dedicated to power for moderator effects in CRTs (e.g., Bloom & Spybrook, 2017; Dong et al., 2018; Spybrook et al., 2016). For example, researchers are considering strategies to improve power to detect moderator effects that involve things such as selecting schools that are more heterogeneous on the moderator of interest to maximize the magnitude of the effect and hence the likelihood of detecting the effect (Zhang et al., 2019). This work extends beyond binary moderators to consider continuous moderators and the potential differences in power across the two types of moderators. In addition, work on power for moderator effects in multisite CRTs is underway. We have also started to see a new set of user-friendly software available to conduct these calculations. PowerUp!-Moderator (Dong, Kelcey, et al., 2016) is available in an excel platform and as *PowerUpR* Shiny application (poweruprshiny.shinyapps.io/v104/). Both of these interfaces are accessible and intended for substantive researchers and methodologists conducting power calculations. We anticipate that the availability of user-friendly tools will ultimately lead to a steady increase in power analyses for planning CRTs to answer questions about *what works, for whom*, and *under what conditions*.

APPENDIX

List of Studies in the Sample

Principal investigator	Project title
Babinski, L.	Efficacy of the DCCS Program: ESL and Classroom Teachers Working Together With Students and Families
Bailey, C.	Promoting School Readiness Through Emotional Intelligence: An Efficacy Trial of Preschool RULER
Bradshaw, C.	Testing the Efficacy of Double Check: A Cultural Proficiency Professional Development Model in Middle Schools
Bradshaw, C	Evaluating Maryland State Policies to Improve School Climate
Bradshaw, C.	Testing the Efficacy of a Developmentally Informed Coping Power Program in Middle Schools
Brown, J.	Testing the Integration of an Empirically-Supported Teacher Consultation Model and a Social-Emotional Learning and Literacy Intervention in Urban Elementary Schools
Bruns, E.	Efficacy of a Brief Intervention Strategy for School Mental Health Clinicians
Crawford, A.	Examining the Cost-effectiveness of Continuous Improvement Models for Preschool Teachers: Balancing PD Structures to Match Teacher Need
Davenport, J.	Improving Children's Understanding of Mathematical Equivalence: An Efficacy Study
Downer, J.	Examining the Efficacy of RULER on School Climate, Teacher Well-Being, Classroom Climate, and Student Outcomes
Dynarski, S.	Dual-Credit Courses and the Road to College: Experimental Evidence From Tennessee

(continued)

APPENDIX. (continued)

Principal investigator	Project title
Feng, M.	Efficacy of ASSISTments Online Homework Support for Middle School Mathematics Learning: A Replication Study
Freiberg, J.	Consistency Management & Cooperative Discipline (CMCD): An Efficacy Trial With Students in Third and Fourth Grade Urban Schools
Gray, S.	Efficacy of the TELL Curriculum for Preschool Children Who Are Economically Disadvantaged
Greenwood, C.	The Effects of Promoting Engaging Early Literacy Interactions in Preschool Environments: Literacy 3D
Gunn, B.	An Investigation of Direct Instruction Spoken English for At-Risk English Learners
Harris, C.	Efficacy Study of an Integrated Science and Literacy Curriculum for Young Learners
Herman, K.	Evaluation of a Classroom Management Training Program for Middle School Teachers
Howard, E.	An Efficacy Trial of the HighScope Preschool Curriculum (HSPC)
Justice, L.	Causal Effects of the Kindergarten Transition Intervention
Landry, S.	Internet Implementation of Empirically-Supported Interventions That Can Be Remotely Delivered in Authentic Preschool Programs for Mothers and Teachers: Evaluation of Direct Child and Teacher Outcomes
Landry, S.	Scalable Approaches for Preparing Early Childhood Teachers: Identifying Costs and Effectiveness of Evidence Based Approaches to Coaching
Lewis, C.	Improvement of Elementary Fractions Instruction: Randomized Controlled Trial Using Lesson Study With a Fractions Resource Kit
Lorch, E.	Efficacy of a Narrative Comprehension Intervention for Elementary School Children At-Risk for Attention-Deficit Hyperactivity Disorder
Mashburn, A.	Efficacy of MindUP on Pre-Kindergarteners' Development of Social-Emotional Learning Competencies and Academic Skills
Moeller, B.	Math for All: Assessing the Efficacy of a Professional Development Program for Elementary School Teachers
Nugent, G.	Testing the Efficacy of INSIGHTS for Promoting Positive Learning Environments and Academic Achievement in Nebraska: A Replication Study
Redmond, C	Testing the Efficacy of Embedded Social Skills Within a Universal Classroom Management Program: Well-Managed Schools
Rosanbalm, M.	Effects of the Incredible Years Dinosaur Classroom Prevention Program on Preschool Children's Executive Functioning and Academic Achievement
Roschelle, J.	Efficacy of an Integrated Digital Elementary School Mathematics Curriculum
Schneider, S.	Word Learning Strategies: A Program for Upper-Elementary Readers
Schneider, S.	Efficacy Study of Adventures Aboard the S.S.GRIN: Social, Emotional, and Academic Skills
Sorby, S.	Enhancing Middle School Mathematics Achievement Through Spatial Skills Instruction
Swanson, E.	Examining the Efficacy of Differential Levels of Professional Development for Teaching Content Area Reading Strategies
Upshur, C.	Kidsteps II: Promoting School Readiness Through Social-Emotional Skill Building in Preschool
Wayne, W.	My Science Tutor: Improving Science Learning Through Tutorial Dialogs (MyST)
Wendt, A.	Evaluation of We Have Skills, A Multimedia Classroom Level Social Skills Program for Elementary Students

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 article is funded by a grant from the National Science Foundation (#1437692) and by a grant from the Institute of Education Sciences (R324U180001).

Note

1. In two studies, we were not able to determine the sample sizes from the structured abstract but were able to find them in associated publications.

References

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547–556.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating parameters of cross-site program effect distributions. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902.

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

Cook, T. D., & Foray, D. (2007). Building the capacity to experiment in schools: A case study of the Institute of Educational Sciences in the US Department of Education. *Economics of Innovation and New Technology*, *16*(5), 385–402.

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *Journal of Experimental Education*, *86*(3), 489–514.

Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2016). *PowerUp!-Moderator: A tool for calculating statistical power and minimum detectable effect size differences of the moderator effects in cluster randomized trials* [Software]. http://www.causalevaluation.org/

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, *40*(4), 334–377.

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

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445–489.

Hedges, L. V., & Rhoads, C. (2009). *Statistical power analysis in education research* (NCSER 2010-3006). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Hill, C., Bloom, H. S., Rebeck-Black, A., & Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Institute of Education Sciences. (2006). Request for applications: Field year 2017. CFDA Number: 84.305A.

Institute of Education Sciences. (2011). Request for applications: Field year 2012. CFDA Number: 84.305A.

Institute of Education Sciences. (2016). Request for applications: Field year 2017. CFDA Number: 84.305A.

Jacob, R., Zhu, P., & Bloom, H. S. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness*, *3*(2), 157–198.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Online Variance Almanac. (n.d.). http://stateva.ci.northwestern.edu/

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

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

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605–627.

Spybrook, J., Lininger, M., & Cullen, A. (2013). From planning to implementation: An examination of changes in the research design, sample size, and statistical power of group randomized trials launched by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness*, *64*(4), 396–342.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *313*(3), 298–318.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education*, *39*(3), 255–267.

Spybrook, J., Westine, C., & Taylor, C. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1), 1–15.

Westine, C., Spybrook, J., & Taylor, J. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519.

Xu, Z. A., & Nichols, A. (2010). *New estimates of design parameters for clustered randomization studies: Findings from North Carolina and Florida* (Working Paper No. 43). National Center for Analysis of Longitudinal Data in Education Research.

Zhang, Q., Spybrook, J., & Tipton, E. (2019, March 8). *The influence of imbalanced subgroup units in statistical power for moderator effects in cluster randomized trials: Empirical evidence from IES-funded studies* [Paper presentation]. Annual Meeting for the Society for Research on Educational Effectiveness, Washington, DC.

Zhu, P., Jacob, R., Bloom, H. S., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Education Evaluation and Policy Analysis*, *34*(1), 45–68.

Authors

JESSACA SPYBROOK is a professor of evaluation, measurement and research at Western Michigan University. Her research focuses on improving the design and analysis of large-scale impact studies in education.

QI ZHANG is a research assistant at Western Michigan University. His research focuses on the design and analysis of experiments, power analyses for multilevel trials, and meta-analysis in education.

BEN KELCEY is an associate professor of quantitative research methodologies in the School of Education at the University of Cincinnati. His research interests include causal inference and measurement methods within the context of multilevel and multidimensional settings such as classrooms and schools.

NIANBO DONG is an associate professor in the School of Education at The University of North Carolina at Chapel Hill. His research interests include causal inference, statistical power analysis, multilevel modeling, and program and policy evaluation.

Manuscript received June 26, 2019
First revision received October 25, 2019
Second revision received April 13, 2020
Accepted April 26, 2020