Experimental Design and Statistical Power for Cluster Randomized Cost-Effectiveness Trials

Wei Li[a], Nianbo Dong[b], Rebecca Maynarad[c], Jessaca Spybrook[d], and Ben Kelcey[e]

[a] College of Education, University of Florida, Gainesville, USA; [b]School of Education, University of North Carolina – Chapel Hill, Chapel Hill, USA; [c]Graduate School of Education, University of Pennsylvania, Philadelphia, USA; [d]College of Education and Human Development, Western Michigan University, Kalamazoo, USA; [e]College of Education, Criminal Justice, and Human Services, University of Cincinnati, Cincinnati, USA

Correspondence concerning this article should be addressed to Wei Li, University of Florida 2711R Norman Hall, Gainesville, FL 32611. Email: wei.li@coe.ufl.edu

Funding:

This article is based on work funded by the National Science Foundation [DRL-2000705]. The opinions expressed herein are those of the authors and not the funding agency.

Citation:

Li, W., Dong, N., Maynard, R. A., Spybrook, J., & Kelcey, B. (2022). Experimental Design and Statistical Power for Cluster Randomized Cost-Effectiveness Trials. *Journal of Research on Educational Effectiveness*. Advance online publication. doi: 10.1080/19345747.2022.2142177

Abstract

Cluster randomized trials (CRTs) are commonly used to evaluate educational interventions, particularly their effectiveness. Recently there has been greater emphasis on using these trials to explore cost-effectiveness. However, methods for establishing the power of cluster randomized cost-effectiveness trials (CRCETs) are limited. This study develops power computation formulas and statistical software to help researchers plan two- and three-level CRCETs. We illustrate the application of our formulas and software for the designs of CRCETs and discuss the influence of sample size, nesting effects, covariates, and the covariance between cost and effectiveness measures on the statistical power of cost-effectiveness estimates.

Keywords: cost-effectiveness analysis, statistical power, cluster randomized cost-effectiveness trials, multilevel models

Experimental Design and Statistical Power for Cluster Randomized Cost-Effectiveness Trials

The randomized controlled trial (RCT) has been seen as the gold standard for evaluating the causal effects of programs, policies, and practices (hereafter referred to as interventions; Imbens & Rubin, 2015). Education interventions often involve nested data structures (e.g., students nested with schools) and, as a result, experiments frequently randomly assign clusters (e.g., schools) rather than individuals to a treatment or control condition (e.g., Conroy et al., 2018; Konstantopoulos et al., 2016; Spybrook et al., 2020). Historically, educational researchers utilized cluster randomized trials (CRTs) to assess the effectiveness of educational interventions but ignored the cost of implementing these interventions (e.g., Bulus & Dong, 2021; Harris, 2009; Shen & Kelcey, 2020). However, policymakers and administrators commonly strive to identify interventions that have maximal effectiveness for a given budget or aim to achieve a target improvement in effectiveness at the lowest possible cost (Levin et al., 2017). Therefore, recent discussions regarding economic evaluations in education call for evaluating the cost as well as the effectiveness of educational interventions to facilitate better decision-making (e.g., Belfield & Bowden, 2019; Levin & Belfield, 2015; Shand & Bowden, 2021).

Evaluations without a credible cost analysis can lead to misleading judgments regarding the relative benefits of alternative strategies for achieving a particular goal (e.g., maximizing the outcomes at current levels of expenditures or minimizing costs for achieving specific effects). For example, results from the Project STAR (Student-Teacher Achievement Ratio) in Tennessee provided strong evidence that class size reduction (CSR) improved student achievement in early grades (Krueger, 1999; Mosteller, 1997; Konstantopoulos & Li, 2012). However, CSR policy is costly (Brewer et al., 1999). Holding constant the level of performance gain, Levin et al. (1987)

estimated that CSR would be more costly than alternative strategies, such as peer tutoring, for improving achievement. Therefore, even though CSR is an effective way to improve student achievement, it might not be the most cost-efficient strategy, or it may simply be not feasible given current resource constraints. Indeed, several European countries stopped implementing CSR policies due to budget limitations (Li & Konstantopoulos, 2016, 2017a).

The CSR example illustrates the importance of incorporating cost analyses for interventions that can produce significant positive effects. Even when a proposed new intervention has similar effects to the older ones (e.g., a null effect), educational researchers and policymakers still need to compare the cost of implementing these interventions for a comprehensive assessment and solid decision-making. For example, in a CRT, Fishman et al. (2013) found no significant difference in the effectiveness of two modalities (i.e., online and face-to-face) of teacher professional development on either teacher or student learning. Still, schools, districts, and states may see online teacher professional development as an attractive alternative to traditional face-to-face professional development programs since online professional development can potentially produce similar effects at much lower costs (e.g., Lay et al., 2020). To sum up, when comparing alternative interventions with similar goals, both the effectiveness and cost should be accounted for if the intent is for the evaluations to support decision-making (Harris, 2009; Levin et al., 2017).

Two primary economic evaluation approaches in education are cost analysis (CA) and cost-effectiveness analysis (CEA). CA identifies all the resources needed to deliver an intervention (Levin & Belfield, 2015). It computes the total cost of an intervention and the average cost per participant and thus can help practitioners and policymakers understand the resources required to replicate a particular intervention. CEA examines the resources needed to

achieve a specific intervention effect and is widely used to compare the cost and effectiveness of different interventions with similar goals (Levin & Belfield, 2015). It provides policymakers and practitioners with estimates of the absolute and relative effectiveness per dollar expended to achieve a specific goal, and thus enables them to determine which intervention alternatives are expected to produce the best outcomes for a given budget cap - a common situation faced by school administrators facing shrinking budgets following the 2008 recession (Sparks, 2019). Today, major educational funding agencies, including the Institute of Education Sciences (IES), are requiring an economic evaluation (e.g., CEA) as part of grant proposals for program evaluations (IES, 2020). Education evaluators are increasingly likely to incorporate cost studies and cost-effectiveness evaluations into their research plans (e.g., Jacob et al., 2016; Mustafa, 2018; Steele et al., 2018).

The study design that involves the random assignment of entire clusters to a treatment or control condition to evaluate both the cost and effectiveness of an intervention is commonly referred to as the cluster randomized cost-effectiveness trial (CRCET). CRCETs link the cost of implementing an intervention to its effect and thus help researchers and policymakers adjudicate the degree to which an intervention is cost-effective. Just as we have become accustomed to designing CRTs with sufficient power, it is crucial to ensure that the size and allocation of the study sample across and within clusters guarantee adequate power (e.g., power > 0.80) to determine whether an intervention is significantly cost-effective or not. This study aims to develop statistical methods and a user-friendly tool to help educational researchers plan their CRCETs.

In education, the incremental cost-effectiveness ratio (ICER), defined as the net cost of an intervention divided by the intervention effect, has been widely used as a cost-effectiveness

measure to compare alternative interventions with similar goals. The intervention with the smallest ICER is deemed the most cost-effective, assuming a positive impact. The current empirical CEA studies in education focus on the descriptive measure of ICER, and do not provide inferential statistics such as *p*-values partially because of the difficulty of conducting statistical inference for ratio statistics (e.g., Bowden & Belfield, 2015; Hollands et al., 2013; Levin et al., 2017; Levin et al., 2012). Without accounting for estimation uncertainty in ICERs, it is difficult to assess whether an intervention is statistically significantly cost-effective. Besides ICER, another commonly used measure is the incremental net monetary benefit (INMB). Because it can facilitate statistical inference and power computation, we used the INMB to measure the cost-effectiveness of an intervention in this study. We provide a more detailed discussion regarding these two measures in the method section.

CRCETs require plans to collect both effectiveness data (e.g., test scores) and cost data in the design phase. Educational researchers commonly utilize the "ingredients method" for cost data collection (Levin et al., 2017), which considers all the ingredients needed to implement an intervention and computes the total cost and average cost per participant of an intervention based on the quantities and prices of the ingredients (Levin et al., 2017). Given the nature of schools, with students nested within classes nested within schools, the ingredients method is commonly applied in the context of a multilevel data structure. Some ingredients are measured at the school level (e.g., school staff and facilities), while others may be measured at the class level (e.g., teacher time) and student level (e.g., volunteer time and transportation for one-to-one tutoring). Student-level cost is computed as the total cost of the student-level ingredients and the average cost per student of the class- and school-level ingredients.

The cost of school-level ingredients usually varies across schools. That is, although the quantities and prices of school-level ingredients are often fixed for students within the same school, they are likely to vary across schools (e.g., Bowden & Belfield, 2015). For example, the renting price for school facilities generally varies across school districts. Similarly, the costs of class- and student-level ingredients (e.g., teacher salaries, student transportation, etc.) may vary within and across schools. As a result, the student-level cost that considers the ingredients at student, class, and school levels varies among individuals and schools. And thus, the school-level cost, computed as the school-level average cost per student, varies among schools. Educational researchers have documented vast variability in ingredients use and cost across schools for a wide range of education interventions ranging from early literacy to college enrollment (Bowden & Belfield, 2015; Hollands et al., 2013; Levin et al., 2017; Levin et al., 2012). For example, Bowden and Belfield (2015) evaluated the cost-effectiveness of the Talent Search program that was created to improve high school completion and college enrollment for disadvantaged students. They found that the school-level costs varied across all categories of ingredients. The school-level average cost per student ranged from $420 to $720, with a standard deviation of 93.9.

It also should be noted that because students are nested within schools and share the same class- and school-level ingredients, student-level costs are correlated among students within the same schools. Similar to effectiveness measures (e.g., test scores), the within-class and school correlations can be represented by the cost data intra-class correlation coefficients (ICCs) at corresponding levels. The nested structure of cost data and the cost variation within levels of measurement should be accounted for in the design of CRCETs. Otherwise, the statistical power and sample size will be misestimated (Li et al., 2020). Compared to traditional impact studies

that only focus on the effectiveness measures, the outcome of interest for CRCETs considers both the effectiveness and cost measures (e.g., INMB). The hierarchical linear model (HLM; Raudenbush & Bryk, 2002) can be used to account for the nested data structure of these measures. When designing a CRCET, educational researchers need to identify the appropriate levels of clustering (e.g., two levels or three levels) and account for the variations of both effectiveness and cost measures when they compute statistical power and sample size (e.g., Li et al., 2020; Manju et al., 2014).

Power analysis methods for effectiveness studies have been widely discussed in education literature (e.g., Dong et al., 2018, 2021; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Kelcey et al., 2019; Konstantopoulos, 2008a, 2008b; Li & Konstantopoulos, 2017b, 2019; Raudenbush, 1997; Raudenbush & Liu, 2000; Schochet, 2008; Spybrook et al., 2011). However, the education literature largely ignores the statistical power of the cost estimates of implementing the interventions in question. Methods for conducting power analysis for randomized cost-effectiveness trials have been discussed in other disciplines (e.g., Willan & Briggs, 2006). For example, health researchers have developed formulas to calculate power for two-level randomized cost-effectiveness trials, where the treatment is at the patient-level or the health care provider-level (Manju et al., 2014, 2015). These methods apply to unconditional models (i.e., no covariates), whereas it is common in education evaluations to include covariates to improve the precision of impact estimates (e.g., Bloom et al., 2007; Hedges & Hedberg, 2007; Konstantopoulos, 2012). These methods developed for the health research field also accommodate nested cost data structure and require individual-level cost data, while it is recommended practice in education research to differentiate costs incurred at different levels of intervention (i.e., individual versus school). Educational researchers commonly collect cost data

through structured interviews with school staff, research partners, and other supplemental services staff for all ingredients, while often missing detail at the student level because of the greater cost of data collection (e.g., Jacob et al., 2016). As a result, student-level cost data are commonly unavailable, and only school-level average costs of student-level ingredients are available (Levin & Belfield, 2015).

Recently, Li et al. (2020) extended power analysis methods used in health science to accommodate the use of covariates in two-level multisite randomized cost-effectiveness trials, where randomization of treatment occurs at the individual level within sites. That same paper discusses the implications of not having student-level cost data for estimating the power. However, that paper did not extend to addressing power computation methods and tools for two- or three-level CRCETs with covariate adjustments, and it did not tackle the issues that arise from the fact that educational interventions commonly involve complicated nesting structures (e.g., students nested with classes, and classes nested within schools).

This study extends the power analysis and CEA literature in three ways. First, we extend the existing power analysis methods for two-level CRCETs (e.g., Manju et al., 2014) to incorporate covariate effects. Second, we develop power analysis methods for three-level CRCETs that also consider covariate effects. Third, we discuss the implications of not having student-level or class-level cost data on statistical power and provide an accessible and user-friendly software program, PowerUp!-CEA, to facilitate planning adequately powered CRCETs. This includes providing practical guidance and illustrative examples regarding how to choose design parameters under various design scenarios and assumptions. The study and accompanying software provide a practical way of designing educational CEA studies to optimize power subject to sample size and allocation constraints.

We begin by presenting a framework for analyzing cluster designs for cost-effectiveness studies that use HLMs. We derive formulas for computing the statistical power to detect a desired effect size regarding the cost-effectiveness of an intervention given sample size and the minimum detectable effect size (MDES; Bloom, 1995) given statistical power and sample size under different scenarios. We then demonstrate the application of the power analysis formulas and discuss the features of statistical power under prototypical assumptions. We conclude with suggestions for extending this work to include different and more complex study designs.

Method

In this section, we first discuss the measures to evaluate the cost-effectiveness of an intervention, and then develop the statistical power formulas for two-level cluster designs. Because the derivation is essentially the same, results for three-level designs are presented in Online Appendix B.

Cost-Effectiveness Measures

In education literature, researchers commonly use the incremental cost-effectiveness ratio (ICER) to measure the cost-effectiveness of an intervention, which is defined as the incremental cost (denoted as ΔC) divided by the incremental effect (denoted as ΔE). The incremental cost is measured by the difference between the average cost for the treatment and control groups. The incremental effect is measured by the difference between the average value of the outcome of interest for those in the treatment and control groups, usually called the average treatment effect (ATE) in experimental studies that focus on the effectiveness measures. Prior CEA studies commonly used the ICER to compare the relative cost-effectiveness among alternative interventions with similar goals (e.g., Bowden & Belfield, 2015; Hollands et al., 2013; Levin et al., 2017; Levin et al. al., 2012). When both the incremental effect and the incremental cost are

positive, the intervention with the smallest ICER is the most cost-effective, and the intervention with an ICER smaller than a targeted value is cost-effective. However, if either the incremental effect or the incremental cost is negative, it is complicated to decide which is the most cost-effective. For example, an intervention with a positive incremental effect but a negative incremental cost might have the same ICER as another intervention with a negative incremental effect but a positive incremental cost; however, the former intervention is more cost-effective than the latter.

Another measure is the incremental net monetary benefit (INMB), defined as

$$INMB = \kappa \Delta E - \Delta C, \quad (1)$$

where κ is the value decision-makers (e.g., society, policymakers, and/or intervention participants) assign to a unit change in the outcome—sometimes referred to as their "willingness-to-pay" (Willan & Briggs, 2006). WTP (or κ) can be interpreted as the threshold ICER that renders the intervention cost-effective (Stinnett & Mullahy, 1998). There are various ways to define κ. For example, Herrera-Araujo et al. (2017) estimated adult's willingness-to-pay to improve reading and speaking skills among dyslexic individuals using a state-preference survey. More generally, when comparing alternative programs with similar outcomes, κ is assumed to be constant and exogenous to the intervention.

Compared to the ICER, the interpretation of an estimated INMB is more straightforward: interventions with a positive INMB would always be deemed cost-effective. Moreover, when the estimated ΔE and ΔC are unbiased estimates, the estimated INMB is also an unbiased estimate of the true INMB, while the estimated ICER is not an unbiased estimator of the true ICER due to it being a ratio estimator (Stinnett & Mullahy, 1998). Therefore, when the sample size is small, the bias in the estimated ICER might not be negligible (Stinnett & Mullahy, 1998). In addition,

statistical inference and power computation are far more straightforward for INMB than for ICER because INMB is a linear function of ΔE and ΔC, while ICER is a ratio of ΔE and ΔC. Specifically, we can easily compute the variance of INMB through

$$Var(INMB) = Var(\kappa\Delta E - \Delta C) = \kappa^2 Var(\Delta E) + Var(\Delta C) - 2\kappa Cov(\Delta E,\ \Delta C). \quad (2)$$

Although it is possible to use Fieller's Theorem to compute the confidence interval for ICER (Willan & O'Brien, 1996), educational evaluators rarely used it in practice. Moreover, the power estimates based on equation (1) or Fieller's Theorem are almost identical, although the underlying assumptions about the distribution of the effectiveness and cost measures differ (see Li et al., 2020). Therefore, to facilitate statistical inference and power computation, as in prior studies (e.g., Manju et al., 2014, 2015; Li et al., 2020), we use the INMB as the CEA measure.

Two-level Cluster Designs: Unconditional Model

Consider a simple two-level cluster randomized design where level-2 units (e.g., schools) are randomly assigned to treatment or control conditions, and the treatment is at the second level. When level-1 cost data are available, two-level HLM can be used to estimate the incremental effect (i.e., ΔE) and the incremental cost (i.e., ΔC), namely

$$E_{ij} = \gamma_{00}^{e} + T_j \Delta E + r_{0j}^{e} + \varepsilon_{ij}^{e}, \quad (3)$$

$$C_{ij} = \gamma_{00}^{c} + T_j \Delta C + r_{0j}^{c} + \varepsilon_{ij}^{c}, \quad (4)$$

and

$$\begin{pmatrix} r_{0j}^{e} \\ r_{0j}^{c} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_e^2 & \tau_{ec} \\ \tau_{ec} & \tau_c^2 \end{pmatrix}\right), \text{ and } \begin{pmatrix} \varepsilon_{ij}^{e} \\ \varepsilon_{ij}^{c} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \sigma_{ec} \\ \sigma_{ec} & \sigma_c^2 \end{pmatrix}\right), \quad (5)$$

where E_{ij} is the effectiveness measure (e.g., test scores) and C_{ij} is the cost for level-1 unit *i* in level-2 unit *j*; T_j is a binary treatment indicator variable; r_{0j}^{e} and r_{0j}^{c} are the level-2 random effects for effectiveness and cost data, respectively; and ε_{ij}^{e} and ε_{ij}^{c} are the level-1 errors for

effectiveness and cost data, respectively. We assume the random effects and level-1 error terms follow bivariate normal distributions as shown in equation (5).

Equations (3) and (4) express the effectiveness measure (e.g., test score) and cost measures on the original scales (e.g., points and dollars), respectively. Following the net-benefit framework (e.g., Manju et al., 2014), we can reconstruct equations (3) and (4) as

$$NMB_{ij} = \kappa E_{ij} - C_{ij} = \pi_{00} + \pi_{01} T_j + r_{0j} + \varepsilon_{ij}, \quad (6)$$

where NMB_{ij} represents the net monetary benefit (NMB) for level-1 unit *i* in level-2 unit *j,* κ is a positive constant that represents the dollar value of willingness-to-pay (Manju et al., 2014; Willan & Briggs, 2006), $\pi_{00} = \kappa\gamma_{00}^{e} - \gamma_{00}^{c}$, $\pi_{01} = \kappa\Delta E - \Delta C$, $r_{0j} = \kappa r_{0j}^{e} - r_{0j}^{c}$, $\varepsilon_{ij} = \kappa\varepsilon_{ij}^{e} - \varepsilon_{ij}^{c}$, $r_{0j} \sim N(0, \tau^2)$, and $\varepsilon_{\text{ij}} \sim N(0, \sigma^2)$. The parameter of interest, π_{01}, represents the INMB of the treatment. When $\pi_{01} > 0$, it indicates the treatment is cost-effective; when $\pi_{01} < 0$, it suggests the treatment is not cost-effective.

Suppose there are *J* level-2 units and *n* level-1 units within each level-2 unit, and thus, the total number of level-1 units is *nJ*. Also, suppose there are J_T units in the treatment group and J_C in the control condition. Define $P = \frac{J_T}{J}$, is the proportion of level-2 units in the treatment group, then the variance of $\hat{\pi}_{01}$ is (Hedges & Rhoads, 2010; Raudenbush, 1997)

$$Var(\hat{\pi}_{01}) = \frac{1}{P(1-P)nJ}(n\tau^2 + \sigma^2). \quad (7)$$

where $\tau^2 = \kappa^2\tau_e^2 + \tau_c^2 - 2\kappa\tau_{ec}$, and $\sigma^2 = \kappa^2\sigma_e^2 + \sigma_c^2 - 2\kappa\sigma_{ec}$. Under the bivariate normal assumptions (i.e., equation 3), the test statistic $t = \frac{\hat{\pi}_{01}}{\sqrt{var(\hat{\pi}_{01})}}$ follows a student's *t* distribution when the null hypothesis is true. When the alternative hypothesis is true, the *t* statistic follows a non-central *t*-distribution with the non-centrality parameter as (Hedges & Rhoads, 2010; Raudenbush, 1997)

$$\lambda = \pi_{01}\sqrt{\frac{P(1-P)nJ}{n\tau^2+\sigma^2}}. \quad (8)$$

Let $\psi_e = \sigma_e^2 + \tau_e^2$ and $\psi_c = \sigma_c^2 + \tau_c^2$ represent the total variance of effectiveness and cost measures, respectively. Prior studies (e.g., Belfield & Bowden, 2019) suggested that the power computation for CEA should allow for identifying a valid impact in dollars. Therefore, we first standardize the effectiveness measures (i.e., $\psi_e = 1$), and then define the effect size as $\delta = INMB = \kappa\Delta E - \Delta C$, representing the net monetary benefit of the intervention with one SD increase of the effective measures. The non-centrality parameter becomes

$$\lambda = \delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(n-1)\rho_e]+\psi_c[(n-1)\rho_c]+(\kappa^2+\psi_c)-2\kappa\sqrt{\psi_c}(nr_2+r_1)}}, \quad (9)$$

where ψ_c can be interpreted as the ratio between ψ_c and ψ_e, given $\psi_e = 1$; $\rho_e = \frac{\tau_e^2}{\psi_e}$ and $\rho_c = \frac{\tau_c^2}{\psi_c}$ are the intra-class correlations (ICCs) of effectiveness and cost data, respectively; and $r_1 = \frac{\sigma_{ec}}{\sqrt{\psi_e\psi_c}}$ and $r_2 = \frac{\tau_{ec}}{\sqrt{\psi_e\psi_c}}$ are the standardized covariance between cost and effectiveness measures at level 1 and level 2, respectively. Note that $r_1 = (1-\rho_e)(1-\rho_c)corr_1$ and $r_2 = \rho_e\rho_c corr_2$ are positively correlated with the correlations between cost and effectiveness measures, where $corr_1 = \frac{\sigma_{ec}}{\sigma_e\sigma_c}$ and $corr_2 = \frac{\tau_{ec}}{\tau_e\tau_c}$ are the correlation coefficients at level 1 and level 2, respectively. Under these specifications, power is defined as

$$Power = 1 - \mathrm{H}\,[c(\alpha/2, J-2), (J-2), \lambda] + \mathrm{H}\,[-c(\alpha/2, J-2), (J-2), \lambda], \quad (10)$$

where $c(\alpha,\nu)$ is the one-tailed critical value of the *t*-distribution with Type I error rate of α and ν degrees of freedom (e.g., $c(0.05,20)=1.72$), $\mathrm{H}(x, \nu, \lambda)$ is the cumulative distribution function of the non-central *t*-distribution with ν degrees of freedom and non-centrality parameter λ. Besides the statistical power of a CRCET, applied researchers also want to estimate the MDES that a

CRCET can detect with sufficient power (e.g., power > 0.8) given sample sizes. The MDES for a two-level cluster design without covariate adjustments is

$$MDES = \frac{M_{J-2}}{\sqrt{P(1-P)nJ}}\sqrt{\kappa^2[(n-1)\rho_e] + \psi_c[(n-1)\rho_c] + (\kappa^2 + \psi_c) - 2\kappa\sqrt{\psi_c}(nr_2 + r_1)}, \quad (11)$$

where $M_{J-2} = t_{\alpha/2} + t_{1-\beta}$ for two-tailed tests with Type I error α, Type II error β, and *J-2* degrees of freedom.

The Online Appendix A provides the derivations of $Var(\widehat{INMB})$ when only cluster level cost information is available, but level-1 effectiveness data are still available. Specifically, we can estimate ΔC through a single-level regression (i.e., equation A1) and estimate ΔE through equation (3). Then based on equation (2) and the results from the Online Appendix A (i.e., equations A3, A4, and A5), we get the variance of $\widehat{INMB}$ as is

$$Var(\widehat{INMB}) = \frac{1}{P(1-P)nJ}[\kappa^2(n\tau_e^2 + \sigma_e^2) + (n\tau_c^2 + \sigma_c^2) - 2\kappa(n\tau_{ec} + \sigma_{ec})]. \quad (12)$$

Again, assume the effectiveness measures are standardized with means of zero and standard deviations of one (i.e., $\psi_e = 1$) and define the effect size as $\delta = INMB$, then the standardized non-centrality parameter is

$$\lambda = \delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(n-1)\rho_e] + \psi_c[(n-1)\rho_c] + (\kappa^2 + \psi_c) - 2\kappa\sqrt{\psi_c}(nr_2 + r_1)}}, \quad (13)$$

which is identical to equation (9). Thus, the power of detecting the cost-effectiveness of treatment is the same for unconditional models regardless of whether or not level-1 cost data are available for two-level CRCETs.

Two-level Cluster Designs: Covariate Effects

When the analysis includes covariates, the two-level HLMs used to estimate the incremental effect (i.e., ΔE) and the incremental cost (i.e., ΔC) become

$$E_{ij} = \gamma_{00}^e + T_j\Delta E + X_{ij}^e\Gamma_{10}^e + Z_j^e\Gamma_{02}^e + r_{A0j}^e + \varepsilon_{Aij}^e, \quad (14)$$

$$C_{ij} = \gamma_{00}^{c} + T_j \Delta C + X_{ij}^{c}\Gamma_{10}^{c} + Z_j^{c}\Gamma_{02}^{c} + r_{A0j}^{c} + \varepsilon_{Aij}^{c}, \quad (15)$$

where X_{ij}^{e} and X_{ij}^{c} are row vectors of level-1 unit characteristics, Γ_{10}^{e} and Γ_{10}^{c} are column vectors of coefficients of level-1 unit characteristics, Z_j^{e} and Z_j^{c} are row vectors of level-2 unit characteristics, and Γ_{02}^{e} and Γ_{02}^{c} are column vectors of coefficients of level-2 unit characteristics. Subscript *A* indicates adjustment because of covariates. The level-1 error terms and the level-2 random effects follow bivariate normal distributions

$$\begin{pmatrix} r_{A0j}^{e} \\ r_{A0j}^{c} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{Re}^2 & \tau_{Rec} \\ \tau_{Rec} & \tau_{Rc}^2 \end{pmatrix}\right), \text{ and } \begin{pmatrix} \varepsilon_{Aij}^{e} \\ \varepsilon_{Aij}^{c} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{Re}^2 & \sigma_{Rec} \\ \sigma_{Rec} & \sigma_{Rc}^2 \end{pmatrix}\right). \quad (16)$$

where subscript *R* indicates residual variance or residual covariance. Then, the NMB for level-1 unit *i* in level-2 unit *j* becomes

$$NMB_{ij} = \pi_{00} + \pi_{A01}T_j + X_{ij}\Gamma_{10} + Z_j\Gamma_{02} + r_{A0j} + \varepsilon_{Aij}. \quad (17)$$

And the non-centrality parameter becomes (Hedges & Rhoades, 2010; Raudenbush, 1997)

$$\lambda = \delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(nw_2^e - w_1^e)\rho_e] + \psi_c[(nw_2^c - w_1^c)\rho_c] + (\kappa^2 w_1^e + \psi_c w_1^c) - 2\kappa\sqrt{\psi_c}(nw_2^{ec}r_2 + w_1^{ec}r_1)}}, \quad (18)$$

where w_1^e and w_2^e represent the unexplained variance of effectiveness data at the first and second levels, respectively; w_1^c and w_2^c represent the unexplained variance of cost data at the first and second levels, respectively; w_1^{ec} and w_2^{ec} represent the unexplained covariance between cost and effectiveness at the first and second levels, respectively. Specifically, $w_1^e = \frac{\sigma_{Re}^2}{\sigma_e^2}$, $w_2^e = \frac{\tau_{RTe}^2}{\tau_{Te}^2}$, $w_1^c = \frac{\sigma_{Rc}^2}{\sigma_c^2}$, $w_2^c = \frac{\tau_{RTc}^2}{\tau_{Tc}^2}$, $w_1^{ec} = \frac{\sigma_{Rec}}{\sigma_{ec}}$, and $w_2^{ec} = \frac{\tau_{RTec}}{\tau_{Tec}}$. Note that we assume group-mean centering of level-1 covariates so that they could only explain a proportion of the variance or covariance at the first level. Then, power is defined as

$$Power = 1 - \mathrm{H}\,[c(\alpha/2, J\text{-}2\text{-}q), (J\text{-}2\text{-}q), \lambda] + \mathrm{H}\,[-c(\alpha/2, J\text{-}2\text{-}q), (J\text{-}2\text{-}q), \lambda], \quad (19)$$

where q is the number of covariates at the second level. All the other terms have been defined previously. Then, the MDES for two-level design with covariate adjustments is

$$MDES(\delta) =$$

$$\frac{M_{J-2-q}}{\sqrt{P(1-P)nJ}}\sqrt{\kappa^2[(nw_2^e - w_1^e)\rho^e] + \psi_c[(nw_2^c - w_1^c)\rho^c] + (\kappa^2 w_1^e + \psi_c w_1^c) - 2\kappa\sqrt{\psi_c}(nw_2^{ec}r_2 + w_1^{ec}r_1)}, \quad (20)$$

where $M_{J\text{-}2\text{-}q} = t_{\alpha/2} + t_{1\text{-}\beta}$ for two-tailed tests with Type I error α, Type II error β, and *J-2-q* degrees of freedom.

When only level-2 cost data are available, but level-1 effectiveness data are available, based on the derivations in the Online Appendix A, we can use a single-level regression to estimate ΔC and use two-level HLM to estimate ΔE, and then get the estimate of INMB (denoted as $\widehat{INMB}$). The variance of $\widehat{INMB}$ becomes

$$Var\left(\widehat{INMB}\right) = \frac{1}{P(1-P)nJ}[\kappa^2(n\tau_{Re}^2 + \sigma_{Re}^2) + (n\tau_{Rc}^2 + \sigma_c^2) - 2\kappa(n\tau_{Rec} + \sigma'_{Rec})]. \quad (21)$$

Again, assume the effectiveness measures are standardized with means of zero and standard deviations of one (i.e., $\psi_e = 1$) and define the standardized effect size as $\delta = INMB$, then the standardized non-centrality parameter becomes

$$\lambda = \delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(nw_2^e - w_1^e)\rho_e] + \psi_c[(nw_2^c - 1)\rho_c] + (\kappa^2 w_1^e + \psi_c) - 2\kappa\sqrt{\psi_c}(nw_2^{ec}r_2 + w'^{ec}_1 r_1)}}, \quad (22)$$

which is not identical to equation (18) because w'^{ec}_1 only takes into account the effects of covariate adjustments in the effectiveness model.

The Online Appendix B provides the derivations and results for three-level cluster designs. Table 1 summarizes standardized non-centrality parameters, MDES, and degrees of freedom for the two- and three-level models we considered in this study. It should be noted that our power and MDES computation methods assume equal cluster sizes (e.g., number of students per school), while in practice it is more likely that the sample sizes vary across clusters. We

suggest using the geometric mean for designs with unbalanced cluster sizes. Because closed-form solutions of power computation formulas are usually unavailable when cluster size varies, prior studies (e.g., Bloom, 2006; Dong et al., 2021; Konstantopoulos, 2010) compared three options – arithmetic mean, harmonic mean, and geometric mean – for power computation. Dong et al. (2021) evaluated the performance of these three options through a simulation analysis and concluded that the power computation using harmonic mean or arithmetic mean either underestimated or overestimates the actual power, while the power calculation based on the geometric mean approximates the actual power from the simulation very well.

Illustration and Discussion

In this section, we first demonstrate the application of our formulas and a free statistical tool (PowerUp!-CEA) to calculate statistical power and MDES and investigate the impact of sample sizes and design parameters on power and MDES estimates for CRCETs, then discuss the similarity and differences between power analyses for CRTs and CRCETs.

Working Example

We illustrate the application of PowerUp-CEA to design CRCETs through a working example, which focuses on the cost-effectiveness of an interim assessment program. Prior studies (e.g., Konstantopoulos et al., 2013) evaluated the impact of Indiana's system of interim assessment on student achievement through a CRT and found significant treatment effects. However, they did not evaluate the cost of implementing this intervention. Assume a new research team plans to design a follow-up whole-school intervention to assess the cost-effectiveness of Indiana's system of interim assessment. Specifically, the research team would like to explore whether the interim assessment is a cost-effective intervention (i.e., $INMB > 0$).

Demonstration

We developed a user-friendly tool-(PowerUp!-CEA) for planning CRCETs that is free and downloadable from the web (https://www.causalevaluation.org/). This tool is implemented through an easy-to-use Microsoft Excel program that channels users to the design and power analysis most appropriate to their study through a dialogue box.

To begin, the research team must first decide the number of total levels within the study. School interventions in education commonly include data and measures at three levels: school-level (e.g., enrollment), class-level (e.g., teacher characteristics), and student-level (e.g., grade level, gender). If they gather data at all three levels (i.e., students, classes, and schools), they can design a three-level CRCET. But, if they lack data at the class level, for example, they would design a two-level CRCET (i.e., students nested within schools). The second step is to decide whether they plan to determine the statistical power they need to achieve given a particular sample and effect size or the MDES achievable for a given sample at a specified level of statistical power. If the research team wants to know the probability of detecting the cost-effectiveness of the intervention with a given effect size, they can use the power calculator within PowerUp!-CEA (see the Online Appendix Table C1 for an example). If the team would like to determine the smallest true effect size that has a given probability (e.g., power > 0.80) of being found to be statistically significant given the specified sample size and allocation, they could use the MDES calculator (see the Online Appendix Table C2 for an example). The third step is to specify the values of design parameters for the effectiveness data, the cost data, the covariance between the effectiveness and cost data, and the statistical significance tests (e.g., alpha level and one/two-tailed test) - the yellow highlighted parameters in the worksheet. Educational researchers typically rely on one of three strategies to estimate these design parameters: (1) calculating them from a pilot study; (2) consulting prior literature for similar

studies; or (3) using existing databases to estimate these parameters (e.g., Spybrook et al., 2016). Once these parameters have been input into PowerUp!-CEA, it automatically calculates the power or MDES.

The relevant design parameters that are common for both CRT and CRCETs include the proportion of variance in the effectiveness measure that is between levels of nesting effects (i.e., ICC), and the proportion of variance in the effectiveness measure that is explained by covariates at different levels. Prior literature (e.g., Hedges & Hedberg, 2007, 2013) suggests that, for a two-level design, a reasonable default estimate of the ICC for student achievement measures (ρ^e) is about 0.23; for a three-level design, reasonable default estimates of ICCs at the second and third levels are 0.08 and 0.15, respectively. Prior literature (e.g., Hedges & Hedberg, 2007, 2013; Konstantopoulos, 2008a), also provides guidance about reasonable default assumptions for the total variance in achievement outcomes explained by covariates at each level (i.e., $w_1^e = w_2^e = w_3^e = 0.5$).

In randomized cost-effectiveness trials, researchers need to make additional assumptions. These include assumptions about the willingness to pay (i.e., κ), the variance and nested effect of cost data, the covariance between cost and effectiveness data, covariate effects, and the availability of student-level and class-level cost data. Regrettably, the compilation of reasonable default assumptions for these measures are still under development. Thus, the study team may need to draw on "guesstimates" to guide their design parameters.

In particular, if the study team plans to collect individual-level cost data and believes the cost will vary among students, classrooms, and schools (i.e., $\psi_c > 0$), study designs should come up with some means of estimating the nesting effects of cost data and the covariance between measures of cost and achievement at each level of the analysis. Manju et al.(2014) reported an

ICC of 0.17 for cost data in a two-level cluster design for a medical study, and they found a negative correlation between the cost and effectiveness measures at the patient-level (level-1) but a positive correlation at the cluster level. For illustration, assume the nesting effects and covariate effects for cost data are similar to those for achievement data (e.g., for two-level designs, $\rho^c = 0.23$ and $w_1^c = w_2^c = 0.5$) in situations where lower-level (e.g., student-level for a two-level design) cost data are available. When lower-level cost data are not available, cost variation and the nesting effects of cost data still impact power and MDES. However, in these cases, it is not possible to use covariates to reduce the cost variance and covariance between cost and effectiveness measures at the corresponding level, therefore, for example, assume $w_1^c = 1$ and $w'^{ec}_1 = 0.8$ for the two-level designs.

The research team can use PowerUp-CEA to compute power and MDES under the above assumptions and three different scenarios regarding the correlations (or covariance) at student- and school- levels for two-level and three-level designs. Tables 2 and 3 summarize the power and MDES estimates. Note that, for three-level designs, the research team assumes the correlations at level 2 are always negative for simplicity. Also, for both tables, the team assumes the effect size is 0.5 (i.e., $\delta_{CRCET} = 0.5$), the total variance of cost data is half as large as the total variance of effectiveness measures (i.e., $\psi_e = 2\psi_c$), a two-sided test with $\alpha = 0.05$, and balanced designs. It also should be noted that almost all prior CEA studies in education ignored $Var(\Delta C)$ as if ΔC was estimated without error, which is equivalent to assume $\psi_c = 0$, as shown in the Online Appendix A. Therefore, Tables 2 and 3 provide power estimates for CRCETs assuming no cost variation to illustrate how power and MDES are incorrectly estimated in that case.

Because the magnitudes of the cost variation and the nesting effects vary across studies, the research team might want to explore the sensitivity of power estimates under different

assumptions, as shown in Figures 2 and 3. These figures illustrate the impact of various design parameters on power using three-level designs as examples. To simplify the presentation and discussion, Figure 2 focuses on unconditional models assuming no correlation between cost and effectiveness measures and evaluates how the cost variation, ICCs of cost data, and the number of level-3 units (e.g., schools) influence power. Specifically, the lines in the left-hand of Figure 2 show how power changes as sample sizes and ICCs of cost data increase, assuming no cost variation; the lines in the middle of Figure 2 show how power changes when the cost measure has a variance twice as large as the effectiveness measure (i.e., $2\psi_e = \psi_c$), and the lines on the right-hand of Figure 2 show how power changes assuming the cost and effectiveness measures have similar variances (i.e., $\psi_e = \psi_c$). Figure 3 incorporates covariate adjustments in the design and displays how covariates influence power under alternative assumptions of cost data availability and the covariance between effectiveness and cost measures. In particular, Unconditional Model (the dotted lines) does not incorporate covariate adjustments at any level; Conditional Model I assumes that only school-level cost data are available, and thus class and student characteristics do not reduce the variance in costs at the student and class levels (i.e., $w_1^c = w_2^c = 1$); and Conditional Model II assumes that student-level and class-level cost data are available and that covariates at both levels reduce the variance of the cost measure and the covariance between cost and effectiveness measures at the first level (e.g., $w_1^c = w_2^c = w_1^{ec} = w_2^{ec} = 0.5$). Figure 3 also assumes that level-1 and level-2 covariates explain less covariance between cost and effectiveness measures when cost data at both levels are missing (e.g., $w_1'^{ec} = w_2'^{ec} = 0.8$), and the class-level standardized covariance are assumed to increase as moving from the left-hand panel (-.03) to the right-hand panel (.06).

It should be noted that examples presented in all Tables and Figures assume equal sample sizes across clusters. For unbalanced cluster sizes (e.g., numbers of classes and students within a school), the researcher has the option of using the geometric mean number of observations per site as a good approximation for the power calculations (Dong et al., 2020). PowerUp!-CEA provides a geometric mean calculator as shown in the Online Appendix Table C6.

Results

The Tables and Figures illustrate how the nested data structure, covariate adjustments, and variation and covariation of cost and effectiveness measures influence power and MDES for the designs of CRCETs. Some results of power analyses for CRCETs are identical or similar to those for traditional power analyses focusing only on estimating the effectiveness of interventions (e.g., a CRT estimating impact on test scores). Specifically, as shown in Figures 2 and 3, power increases as the sample size increases but decreases as the ICC of effectiveness or cost measure increases. Also, incorporating covariates adjustments at each level increases the statistical power if covariates can explain a meaningful proportion of the effectiveness or cost measure variation at the corresponding level. Therefore, collecting class- and student-level cost data can increase power or decrease MDES, as shown in Figure 3.

Tables 2 and 3 and Figures 1 and 2 also reveal some specific and important findings for CRCETs. First, power and MDES are misestimated when ignoring cost variation. Tables 2 and 3 compare the power and MDES estimates for designs that assume no cost variation to those that assume costs vary among students, classrooms, and schools. We can see that for both two- and three-level CRCETs, the power estimates (or MDES estimates) are sensitive to whether cost variation is accounted for and to the direction and strength of the covariance between cost and effectiveness data. For example, as shown in Table 2, row 2, for the same design, the estimated

power is 0.485 for a conditional model that assumes no cost variation. Assuming costs vary at both student and school levels, but there is no correlation between cost and effectiveness data, the power decreases to 0.441. When we assume that costs vary and are positively correlated with the effectiveness measure, the power increases to 0.559, but when they vary and are negatively correlated with the effectiveness measure, the power decreases to 0.365. We find similar results for three-level designs in Table 3. These findings indicate that ignoring cost variation in a CRCET can result in over or underestimation of statistical power (and MDES).

The ratio of the cost variation to the effectiveness variation and the nesting effects of the cost data have a negative impact on power. As shown in Figure 2, other things being equal, power is greater when $\psi_e = \psi_c$ than when variances are unequal (e.g., $2\psi_e = \psi_c$). It also illustrates that other things being equal, power decreases as the ICC of cost data increases. Based on the non-centrality parameter formulas (e.g., equation 8), the effect of a higher ICC depends on the magnitudes of ψ_c: the impact increases as the total variance of cost data increases.

Second, covariance (or correlations) between effectiveness and cost data influences the power and MDES estimates. For example, assume a conditional model with both student level and classroom level cost data available, as shown in the last row of Table 3. When intervention cost and student achievement are positively correlated at both the student and school level, for instance $r_1 = 0.1$, $r_2 = -0.03$, and $r_3 = 0.07$, the MDES is about 0.455. When they are negatively correlated, for instance $r_1 = -0.1$, $r_2 = -0.03$, and $r_3 = -0.07$, the MDES increases to 0.563. And when they are positively correlated at the school level but negatively correlated at the student level (e.g., $r_1 = -0.1$, $r_2 = -0.03$, and $r_3 = 0.07$), the MDES becomes 0.458.

Other things being equal, as the covariance (i.e., r_1, r_2, and r_3) change from negative to positive, power increases monotonically. That is, when cost and effectiveness measures are

positively correlated at any level, power increases as the strength of the correlation (or the absolute value of the standardized covariance) increases. And when the cost and effectiveness measures are negatively correlated, power decreases as the strength of the correlation increases. That is because, as shown in equation (1), the variance of INMB is negatively correlated with the covariance between ΔE and ΔC. Therefore, when the correlation is positive, the absolute value of the standardized covariance is negatively correlated with the variance of INMB. Conversely, when the correlation is negative, the absolute value of the standardized covariance is positively correlated with the variance of INMB. Since power increases as the variance of INMB decreases, when the correlation is positive, the stronger the correlations are, the larger the power is; when the correlation is negative, the stronger the correlations are, the smaller the power is. It also should be noted that the level-3 standardized covariance (or correlation) have a more significant impact on power comparing the correlations at the first level.

Third, covariates influence power and MDES estimates by changing the covariance between cost and effectiveness measures. Figure 3 illustrates this finding by comparing power across unconditional models (the dotted lines) and conditional models that incorporate student-level, class-level, and school-level covariates (the dashed and solid lines). It indicates that when covariates explain a proportion of the covariance between cost and effectiveness measures at a particular level, their inclusion in the analysis affects power. However, the direction of impact depends on the sign of the covariance. Specifically, assuming covariates could decrease the absolute value of the covariance, and based on equation 1, when the correlations are positive, covariates tend to increase the variance of INMB, and, thus, decrease the power and vice versa. As illustrated in Figure 3, when the standardized covariance are negative at both the student- and classroom levels (left-hand lines), power estimates based on Conditional Model II are larger than

those from Conditional Model I. Holding all the other parameters fixed, when the class-level standardized covariance increases from -0.03 to 0.03, the power estimates from the two conditional models (lines in the middle of Figure 3) are almost the same. But, holding all the other parameters fixed, when the class-level standardized covariance increases to 0.06 (right-hand lines), power estimates from Conditional Model I are notably larger than those from Conditional Model II. These findings also indicate that, when the cost and effectiveness measures are negatively correlated at all levels, collecting cost data at lower levels can increase power through covariate adjustments. In contrast, when the correlations are positively correlated at certain levels, the impact of cost data and covariates on power are indeterminate: covariate adjustments can reduce covariate at corresponding levels but also increase the covariance and thus do not always increase power. That is, collecting student-level and class-level cost information does not always help increase power or reduce MDES. It also should be noted that, as discussed in the method section, power and MDES estimates based on unconditional models are not sensitive to the availability of individual-level cost data. The power and MDES estimates for the unconditional models in Tables 2 and 3 illustrate this finding.

Comparisons between Power Analyses for CRTs and CRCETs

There are a couple of key differences in computing power or MDES when the research team designs a CRCET compared to designing a CRT. First, the measure of interest and the study purpose are different. Compared to a CRT that focuses on the effectiveness measure (e.g., test scores), a CRCET focuses on the cost-effectiveness measure (e.g., NMB or INMB) that combines both the effectiveness measure (e.g., test scores) and cost measures (e.g., student-level cost to implement the interim achievement). Therefore, the purpose of a power analysis for a CRT is to determine the minimum required sample sizes to test whether the treatment effect is

larger than zero (i.e., $\Delta E > 0$) with confidence. In contrast, the purpose of a power analysis for a CRCET becomes, for example, to identify the minimum required sample size to test whether the INMB is larger than zero (i.e., INMB > 0) with confidence. Second, the power analysis for a CRT only considers the variance of the effectiveness measure, whereas the power analysis for a CRCET considers both the variance of the effectiveness measure and the cost measure and their covariance. As a result, for the same designs with the same sample sizes at all levels and the same design parameters, the power to detect the cost-effectiveness of the intervention tends to be different than that to detect the effectiveness of the intervention.

For example, assume a two-level cluster design without covariate adjustments and a standardized effectiveness measure with a mean of zero and SD of one (i.e., $\psi_e = 1$). Raudenbush (1997) provided the non-centrality parameter for a t-test to check whether the average treatment effect (i.e., the incremental effect, ΔE) is larger than zero:

$$\lambda_{CRT} = \Delta E \sqrt{\frac{P(1-P)nJ}{(n-1)\rho_e+1}}. \quad (23)$$

First, to make the comparisons of power estimates between a CRT and a CRCET simpler under the same design (i.e., the same sample sizes, design parameters, and ΔE), we assume the research team knows the true population value of the incremental cost, and thus there is no cost variation (i.e., $\psi_c = 0$) and $Var\ (\Delta C) = 0$ based on equation A3 from the Online Appendix A. Therefore, according to equation (13) and assume $\psi_e = 1$, the non-centrality parameter for a *t*-test to check whether an intervention is cost-effective (i.e., $INMB = \kappa\Delta E - \Delta C > 0$) is:

$$\lambda_{CRCET} = INMB \sqrt{\frac{P(1-P)nJ}{\kappa^2[(n-1)\rho_e]+0\times[(n-1)\rho_c]+(\kappa^2+0)-2\kappa\times0\times(nr_2+r_1)}} = (\Delta E - \frac{\Delta C}{\kappa})\sqrt{\frac{P(1-P)nJ}{(n-1)\rho_e+1}}. \quad (24)$$

It should be noted that, an intervention is deemed cost effective if and only if $INMB > 0$ (or $\Delta E > \frac{\Delta C}{\kappa}$). In particular, the average treatment effect (i.e., ΔE) does not necessarily need to be

larger than zero. For example, a cost-saving intervention (i.e., $\Delta C < 0$) with a negative treatment effect (i.e., $\Delta E < 0$) could still be cost-effective. Comparing equations (23) and (24), we can see that $\lambda_{CRT} = \lambda_{CRCET}$ only when $\Delta C = 0$, assuming no cost variation (i.e., $\psi_c = 0$). That is, even if researchers know the population value of ΔC, the power for a CRCET is likely to differ from that for an otherwise similarly designed CRT because ΔC is not equal to zero in general.

Second, educational researchers commonly do not know the population value of ΔC and prior CEA studies in education found substantial cost variation among clusters (e.g., Bowden & Belfield, 2015). Therefore, we recommend considering the cost variation and its nested structure (ψ_c and ρ_c) when designing CRCETs. Then, the differences in power estimates between CRTs and CRCETs depend on the specific values of the design parameters, such as the incremental cost (ΔC), cost data variation (ψ_c) and ICC (ρ_c), covariate effects, the covariance between cost and effectiveness data, etc. For the same design, the power from a CRT tend to be larger than that from a CRCET because of the cost variation (i.e., $\psi_c \neq 0$) and the incremental cost (i.e., $\Delta C \neq 0$). However, the power to detect the cost-effectiveness of an intervention can still be larger than the power to detect the treatment effect for the same design. For example, when the cost and effectiveness data are positively correlated, according to equation 1, the variance of INMB might be smaller than the variance of the incremental effect (i.e., ΔE), and thus the statistical power for a CRCET might be larger than a CRT with the same design parameters and sample sizes. Also, for some interventions that could reduce cost (e.g., an online teacher PD program compared to a face-to-face PD program), λ_{CRCET} is large than λ_{CRT}, and thus the power of a CRCET might be larger than that of a CRT.

To illustrate such a possibility, we consider two scenarios where the incremental cost is either positive (i.e., $\Delta E = 0.4$, $\Delta C = 0.3$, and $\kappa = 2$) or negative (i.e., $\Delta E = 0.2$, $\Delta C = -0.1$,

and $\kappa = 2$) as examples. Assuming the effectiveness measures are standardized (i.e., $\sqrt{\psi_e} = 1$), the effect sizes for the two scenarios are the same: $\delta_1 = \delta_2 = 0.5$. And thus, the power to detect whether the intervention is cost-effectiveness for these two scenarios is the same under the same sample sizes and design parameters; however, the power to detect whether the intervention is effective (e.g., $\Delta E \neq 0$) is not the same because the treatment effects are different (i.e., $\Delta E = 0.4$ or 0.2). The last two columns of Tables 2 and 3 summarize the power or MDES estimates from CRTs for these two scenarios. We can see that, for instance, as shown in Table 2, power estimates for CRCETs are consistently smaller than those from CRTs if the incremental cost is positive. However, when the incremental cost is negative, the power estimates from CRTs are smaller than those from CRCETs. It should be noted that we assumed relatively smaller covariance when they are positive (i.e., $r_1 = r_2 = 0.1$). If the standardized covariance increase to 0.25 (i.e., $r_1 = r2 = 0.25$), the power for a two-level CRCET without covariates adjustment (i.e., an unconditional model) becomes 0.868, which is larger than that from a CRT (0.803). Similar results were also found in Table 3.

Conclusion

CRTs are becoming more common in education to evaluate interventions. Often educational researchers focus on the effectiveness measures (e.g., test scores) but ignore the cost of delivering the intervention. Recent studies (e.g., Belfield & Bowden, 2019; Levin et al., 2017; Shand & Bowden, 2021) suggest educational evaluation should analyze both the cost and effectiveness of an intervention for sound decision-making. CRCETs help researchers examine whether an intervention is cost-effective. Still, the education literature has not previously had well-documented procedures for conducting power analysis to guide the planning of such studies. In this study, we extended previous work on power analysis for CRCETs (e.g., Willan &

Briggs, 2006; Manju et al., 2014, 2015; Li et al., 2020) from two-level to three-level designs, and presented methods for educational researchers who usually incorporate covariate effects and only have cluster-level cost data. In general, the power of the test of the cost-effectiveness of an intervention for CRCETs is a function of the effect size, the sample sizes at each level, the nesting effects of effectiveness and cost data, the covariance between cost and effectiveness at each level, the ratio between the total variance of cost and the total variance of effectiveness, and the proportion of the variances and covariance between cost and effectiveness that covariates explain. We also implemented our formulas to a free tool – PowerUp!-CEA to help researchers plan CRCETs.

Our study has shown that the power and MDES estimates for CRCETs differ from those for CRTs with the same sample sizes and design parameters. When the incremental cost is positive, the power for a CRCET may be smaller than that from a CRT because of a smaller effect size and (or) a larger variance, holding the sample size and design parameters fixed. When the incremental cost is negative, or the covariance between cost and effectiveness are positive and high, the power for a CRCET can be larger than that of a CRT because of a larger effect size or a smaller variance. Educational researchers should conduct a different power for CRCETs to guarantee a good enough chance of detecting a cost-effective intervention. Cost variation and the nested structure of the cost data should be accounted for when performing power analysis for CRCETs, otherwise, the power or MDES may be misestimated.

The methods we developed in this study apply to the design and analysis of CRCETs regardless of whether lower-level cost data (e.g., student- or class-level) are available. For example, as shown in Table 1 and Online Appendices A and B, when level-1 cost data are available, researchers can use HLMs to compute power and INMB; when only cluster-level cost

data are available, researchers can use the single-level method (e.g., OLS regression) to calculate power and INMB if all related design parameters are available. Note that, for models without lower-level cost data, we used the same definitions of design parameters (e.g., effect size, ICC of cost data, etc.) and the same notations as those from HLM to present the cost variation (e.g., equations A2 and B10 in the Online Appendices), which allows us to compare power or MDES for designs with or without lower-level cost information. Based on the results from Tables 2 and 3, we know that collecting lower-level cost data (e.g., student- or class-level) can increase power if the covariates at the corresponding level could explain a meaningful portion of the cost variation at that level. However, it usually requires substantial resources for student- or class-level cost data collection, which can be used to sample more schools to boost power. Therefore, one promising direction of future research is to consider the study budget when computing power or MDES for CRCETs, and to explore which strategy – collecting lower-level cost information or sampling more clusters – can maximize power or minimize the MDES through the optimal design framework.

To use our formulas and tool effectively in planning CRCETs, researchers should be prepared to make informed judgments about the value of the design parameters (e.g., ICCs for cost data, the covariance between cost and effectiveness, etc.). Prior studies have documented empirical values of the design parameters for effectiveness measures (e.g., Dong et al., 2016; Hedberg & Hedges, 2014; Hedges & Hedberg, 2007; Kelcey et al., 2017; Westine et al., 2020); however, to date, there is very limited information regarding the cost data or the covariance between the cost data and the effectiveness data. Therefore, one important direction for further work is the development of empirically-based estimates of these parameters. Considering a growing number of educational interventions (e.g., IES-funded projects) that include an

economic evaluation component, we suggest that they make cost-related information such as cost data ICCs and the covariance between cost and effectiveness measures publicly available, which can inform power analyses for CRCETs. It should be noted that the power analysis methods we developed in this study require empirical estimates of some design parameters related to lower-level cost data (e.g., ρ^c, w_1^c, w'^{ec}_1, and r_1) even for designs that only plan to collect cluster-level cost data. However, these parameters cannot be estimated using cost data from most prior CEA studies in education that did not collect student- or class-level cost information. Future research needs to provide alternative power computation formulas that only require information from aggregated cluster-level cost data.

A limitation of the current study is that we assume the available cost data resulting from the application of the ingredients method will be accurately measured. Yet, the reality is that there are likely to be varying degrees of error in measurement, depending on the type of intervention and the evaluators' ability to access data from various sources. Cox and Kelcey (2019) found that the measurement error of effectiveness measures (e.g., test scores) negatively affects power and MDES. Similarly, holding all the other factors fixed, the power estimate should become smaller when the cost measures are not accurately measured. However, there are no studies either addressing the validity of cost estimates generated using the ingredients method or the reliability (or measurement error) of the cost estimates generated under varying types of interventions and settings or drawing on different information sources. The second direction of future research is exploring how errors measuring costs of interventions affect power and MDES. Finally, in so far as education research commonly includes even more complicated designs than considered here (e.g., treatment at the student or class level), the work should be extended to accommodate three-level multisite designs.

References

Belfield, C. R., & Bowden, B. (2019). Using resource and cost considerations to support educational evaluation: Six domains. *Educational Researcher*, *48*(2), 120–127. https://doi.10.3102/0013189X18814447

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 10(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research* (MDRC Working Papers on Research Methodology). http://www.mdrc.org/publications/437/full.pdf

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29(1)*, 30-59. https://doi.org/10.3102/0162373707299550

Bowden, A., & Belfield, C. (2015). Evaluating the talent search TRIO program: A benefit-cost analysis and cost-effectiveness analysis. *Journal of Benefit-Cost Analysis*, 6(3), 572-602. https://doi.org/10.1017/bca.2015.48

Brewer, D. J., Krop, C., Gill, B. P., & Reichardt, R. (1999). Estimating the costs of national class size reductions under different policy alternatives. *Educational Evaluation and Policy Analysis*, *21*(2), 179–192. https://doi.org/10.2307/1164299

Bulus, M., & Dong, N. (2021). Bound constrained optimization of sample sizes subject to monetary restrictions in planning multilevel randomized trials and regression

discontinuity studies. *Journal of Experimental Education*, *89*(2), 379-401. https://doi.org/10.1080/00220973.2019.1636197

Bulus, M, Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: power analysis tools for multilevel randomized experiments. (Version 1.1.0) [Software]. Available from https://www.r-pkg.org/pkg/PowerUpR

Conroy, M.A**.,** Sutherland, K.S., Algina, J., Werch, B., & Ladwig, C. (2018). Prevention and treatment of problem behaviors in young children: Clinical implications from a randomized controlled trial of BEST in CLASS. *American Education Research Journal, 4,* 1-16. https://doi.10.1177/2332858417750376

Cox, K., & Kelcey, B. (2019). Optimal design of cluster- and multisite-randomized studies using fallible outcome measures. *Evaluation Review*, *43*(3–4), 189–225. https://doi.org/10.1177/0193841X19870878

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *The Journal of Experimental Education*, *86*(3), 489-514. https://doi.org/10.1080/00220973.2017.1315714

Dong, N., Kelcey, B., & Spybrook, J. (2021). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics, 46*(5), 527-559. https://doi.org/10.3102/1076998620961492

Dong, N., & Maynard, R. (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. https://doi.org/10.1080/19345747.2012.673143

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 panning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review, 40*(4), 334-377. https://doi.10.1177/0193841X16671283

Harris, D. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3-29. https://doi.org/10.3102/0162373708327524

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlations values for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.10.3102/0162373707299706

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. https://doi.org/10.1177/0193841X14529126

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review, 38*(6), 546-582. https://doi.org/10.1177/0193841X14554212

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

Herrera-Araujo, D., Shaywitz, B. A., Holahan, J. M., Marchione, K. E., Michaels, R., Shaywitz, S. E., & Hammitt, J. K. (2017). Evaluating willingness to pay as a measure of the impact

of dyslexia in adults. *Journal of Benefit-Cost Analysis*, *8*(1), 24-48. https://doi.10.1017/bca.2017.3

Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H. M., Belfield, C. R., ... & Hanisch-Cerda, B. (2013). Improving early literacy: Cost-effectiveness analysis of effective reading programs. *Center for Benefit-Cost Studies of Education, Teachers College, Columbia University*. http://frg.vkcsites.org/wp-content/uploads/2018/07/KPALS-PDF-Improving-Early-Literacy.pdf

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Institute of Education Sciences (IES). (2020). Request for application: Education research grants. U.S. Department of Education. https://ies.ed.gov/funding/pdf/2021_84305A.pdf

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness*, *9*(sup1), 67-92. https://doi.org/10.1080/19345747.2016.1138560

Kelcey, B., Spybrook, J., & Dong, N. (2019). Sample size planning for cluster-randomized interventions probing multilevel mediation. *Prevention Science*, *20*, 407-418. https://doi.org/10.1007/s11121-018-0921-6

Kelcey, B., Spybrook, J., Phelps, G., Jones, N., & Zhang, J. (2017). Designing large-scale multisite and cluster-randomized studies of professional development. *The Journal of Experimental Education, 85*(3), 389-410. https://doi.org/10.1080/00220973.2016.1220911

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*(1), 66-88. https://doi.org/10.1080/19345740701692522

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265-288. https://doi.org/10.1080/19345740802328216

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *Journal of Experimental Education, 78*(3), 291-317. https://doi.org/10.1080/00220970903292876

Konstantopoulos, S. (2012) The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, *47*(3), 392-420. https://doi.org/10.1080/00273171.2012.673898

Konstantopoulos, S., & Li, W. (2012). Modeling class size effects across the achievement distribution. *International Journal of Sociology of Education*, *1*(1), 5-26. https://doi.org/10.4471/rise.2012.01

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35*(4), 481-499. https://doi.org/10.3102/0162373713498930

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness, 9*(S1), 188-208. https://doi.org/10.1080/19345747.2015.1116031

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*(2), 497-532. https://doi.org/10.1162/003355399556052

Lay, C. D., Allman, B., Cutri, R. M., & Kimmons, R. (2020). Examining a decade of research in online teacher professional development. *Frontiers in Education, 5*, 1–10. https://doi.org/10.3389/feduc.2020.573129

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, *8*(3), 400-418. https://doi.org/10.1080/19345747.2014.915604

Levin, H. M., Belfield, C., Hollands, F., Bowden, A. B., Cheng, H., Shand, R., ... & Hanisch-Cerda, B. (2012). Cost-effectiveness analysis of interventions that improve high school completion. *Teacher College, Columbia University.*

Levin, H. M., Glass, G. V., & Meister, G. (1987). Cost-effectiveness of computer-assisted instruction. *Evaluation Review, 11*(1), 50–72. https://doi.org/10.1177/0193841X8701100103

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2017). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. Sage publications.

Li, W., Dong, N., & Maynard, R. A. (2020). Power analysis for two-level multisite randomized cost-effectiveness trials. *Journal of Educational and Behavioral Statistics*, *45*(6), 690–718. https://doi.org/10.3102/1076998620911916

Li, W., & Konstantopoulos, S. (2016). Class size effects on fourth-grade mathematics achievement: Evidence from TIMSS 2011. *Journal of Research on Educational Effectiveness, 9(4),* 503-530. https://doi.org/10.1080/19345747.2015.1105893

Li, W., & Konstantopoulos, S. (2017a). Does class size reduction close the achievement gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement, 28,* 292-330. https://doi.org/10.1080/09243453.2017.1280062

Li, W., & Konstantopoulos, S. (2017b). Power analysis for models of change in cluster randomized designs. *Educational and Psychological Measurement*, *77*, 119-142. https://doi.org/10.1177/0013164416641460

Li, W., & Konstantopoulos, S. (2019). Power computations for polynomial change in block randomized designs. *Journal of Experimental Education*, *87*(4), 575-595. https://doi.org/10.1080/00220973.2018.1496057

Manju, M., Candel, M. J., & Berger, M. P. (2014). Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Statistics in Medicine*, *33*(15), 2538-2553. https://doi.org/10.1002/sim.6112

Manju, M. A., Candel, M. J., & Berger, M. P. (2015). Optimal and maximin sample sizes for multicentre cost-effectiveness trials. *Statistical Methods in Medical Research, 24*(5), 513-539. https://doi.org/10.1177/0962280215569293

Mosteller, F. (1997). The Tennessee study of class size in the early school grades. *Bulletin of the American Academy of Arts and Sciences, 50*(7), 14-25. https://doi.org/10.2307/3824562

Mustafa, N. (2018). Cost-effectiveness analysis: Educational interventions that reduce the incidence of HIV/AIDS infection in Kenyan teenagers. *International Journal of Educational Development*, *62*, 264-269. https://doi.org/10.1016/j.ijedudev.2018.06.001

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173-185. https://doi.10.1037/1082-989X.2.2.173

Raudenbush, S. W. & Bryk, A. S. (2002*). Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199-213. https://doi.org/10.1037/1082-989X.5.2.199

Robinson, R. (1993). Cost-effectiveness analysis. *British Medical Journal*, *307*, 924-926. Doi: https://doi.org/10.1136/bmj.307.6909.924

Schochet, P. Z. (2008). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. https://doi.org/10.3102/1076998607302714

Shand, R., & Bowden, A. B. (2021). Empirical support for establishing common assumptions in cost research in education. *Journal of Research on Educational Effectiveness*, 1-27. https://doi.org/10.1080/19345747.2021.1938315

Shen, Z., & Kelcey, B. (2020). Optimal sample allocation under unequal costs in cluster-randomized trials. *Journal of Educational and Behavioral Statistics, 45*(4), 446-474. https://doi.org/10.3102/1076998620912418

Sparks, S. (2019, April 09*). More education studies look at cost-effectiveness*. Education Week. https://www.edweek.org/leadership/more-education-studies-look-at-cost-effectiveness/2019/04.

Spybrook, J., Raudenbush, S. W., Liu, X. F., Congdon, R., & Martínez, A. (2011). *Optimal design for longitudinal and multilevel research: documentation for the optimal design software version 3.0*. www.wtgrantfoundation.org.

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*. https://doi.org/10.1177/2332858415625975

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine What Works, For Whom, and Under What Conditions. *Educational Evaluation and Policy Analysis*, *42*(3), 354–374. https://doi.10.3102/0162373720929018

Steele, J. L., Slater, R. O., Li, J., Zamarro, G., Miller, T., & Bacon, M. (2018). Dual-Language Immersion Education at Scale: An Analysis of Program Costs, Mechanisms, and Moderators. *Educational Evaluation and Policy Analysis*, *40*(3), 420-445. https://doi.org/10.3102/0162373718779457

Stinnett, A., & Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, *18*(2_suppl), S68-S80. https://doi.org/10.1177/0272989X98018002S09

Westine, C. D., Unlu, F., Taylor, J., Spybrook, J., Zhang, Q., & Anderson, B. (2020). Design parameter values for impact evaluations of science and mathematics interventions involving teacher outcomes. *Journal of Research on Educational Effectiveness, 13*(4), 816-839. https://doi.org/10.1080/19345747.2020.1821849

Willan, A. R., & Briggs, A. H. (2006). *Statistical analysis of cost-effectiveness data* (Vol. 37). John Wiley & Sons.Willan, A. R., & O'Brien, B. J. (1996). Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. *Health Economics*, *5*(4), 297-305. https://doi.org/10.1002/(SICI)1099-1050(199607)5:4<297::AID-HEC216>3.0.CO;2-T

Table 1. Summary of the Standardized Noncentrality Parameter, MDES, and Degrees of Freedom

Model Name	HLM	Standardized Noncentrality Parameter (λ) and MDES	Degrees of Freedom
Two-Level Model: Level-1 Cost Data are Available	$E_{ij} = \gamma^e_{00} + \gamma^e_{A01}T_j + X^e_{ij}\Gamma^e_{10} + Z^e_j\Gamma^e_{02} + r^e_{Aoj} + \varepsilon^e_{Aij}$, $C_{ij} = \gamma^c_{00} + \gamma^c_{A01}T_j + X^c_{ij}\Gamma^c_{10} + Z^c_j\Gamma^c_{02} + r^c_{A0j} + \varepsilon^c_{Aij}$.	Standardized Noncentrality Parameter (λ): $\delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(nw^e_2 - w^e_1)\rho_e] + \psi_c[(nw^c_2 - w^c_1)\rho_c] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nw^{ec}_2 r_2 + w^{ec}_1 r_1)}}$ MDES: $\frac{M_{J-2-q}}{\sqrt{P(1-P)nJ}}\sqrt{\kappa^2[(nw^e_2 - w^e_1)\rho^e] + \psi_c[(nw^c_2 - w^c_1)\rho^c] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nw^{ec}_2 r_2 + w^{ec}_1 r_1)}$	*J-2-q*
Two-Level Model: Level-2 Cost Data are Available	$E_{ij} = \gamma^e_{00} + \gamma^e_{A01}T_j + X^e_{ij}\Gamma^e_{10} + Z^e_j\Gamma^e_{02} + r^e_{Aoj} + \varepsilon^e_{Aij}$, $C_j = \gamma^c_{00} + \gamma^c_{01}T_j + r^c_j$.	Standardized Noncentrality Parameter (λ): $\delta\sqrt{\frac{P(1-P)nJ}{\kappa^2[(nw^e_2 - w^e_1)\rho_e] + \psi_c[(nw^c_2 - 1)\rho_c] + (\kappa^2 w^e_1 + \psi_c) - 2\kappa\sqrt{\psi_c}(nw^{ec}_2 r_2 + w'^{ec}_1 r_1)}}$ MDES: $\frac{M_{J-2-q}}{\sqrt{P(1-P)nJ}}\sqrt{\kappa^2[(nw^e_2 - w^e_1)\rho^e] + \psi_c[(nw^c_2 - w^c_1)\rho^c] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nw^{ec}_2 r_2 + w'^{ec}_1 r_1)}$	*J-2-q*
Three-Level Model: Level-1 Cost Data are Available	$e_{ijl} = \gamma^e_{000} + \gamma^e_{A001}T_l + X^e_{ijl}\Gamma^e_{100} + Z^e_{jl}\Gamma^e_{010} + W^e_l\Gamma^e_{002} + u^e_{A00l} + r^e_{A0jl} + \varepsilon^e_{Aijl}$, $c_{ijl} = \gamma^c_{000} + \gamma^c_{A001}T_l + X^c_{ijl}\Gamma^c_{100} + Z^c_{jl}\Gamma^c_{010} + W^c_l\Gamma^c_{002} + u^c_{A00l} + r^c_{A0jl} + \varepsilon^c_{Aijl}$.	Standardized Noncentrality Parameter (λ): $\sqrt{\frac{\delta^2 P(1-P)nJL}{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw^{ec}_2 r_2 + w^{ec}_1 r_1)}}$ MDES: $\frac{M_{L-2-g}}{\sqrt{P(1-P)nJL}} \times$ $\sqrt{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw^{ec}_2 r_2 + w^{ec}_1 r_1)}$	*L-2-g*

Table 1. *(continued)*

Model Name	HLM	Standardized Noncentrality Parameter (λ) and MDES	Degrees of Freedom
Three-Level Model: Level-2 Cost Data are Available	$e_{ijl} = \gamma^e_{000} + \gamma^e_{A001}T_l + X^e_{ijl}\Gamma^e_{100} + Z^e_{jl}\Gamma^e_{010} + W^e_l\Gamma^e_{002}$ $+u^e_{A00l} + r^e_{Aojl} + \varepsilon^e_{Aijl}$, $c_{jl} = \gamma^c_{000} + \gamma^c_{A001}T_l + Z^c_{jl}\Gamma^c_{010} + W^c_l\Gamma^c_{002}$ $+u^c_{A0l} + r^c_{Ajl}$.	Standardized Noncentrality Parameter (λ): $\sqrt{\frac{\delta^2 P(1-P)nJL}{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw^{ec}_2 r_2 + w'^{ec}_1 r_1)}}$ MDES: $\frac{M_{L-2-g}}{\sqrt{P(1-P)nJL}} \times$ $\sqrt{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw^{ec}_2 r_2 + w'^{ec}_1 r_1)}$	*L-2-g*
Three-Level Model: Level-3 Cost Data are Available	$e_{ijl} = \gamma^e_{000} + \gamma^e_{A001}T_l + X^e_{ijl}\Gamma^e_{100} + Z^e_{jl}\Gamma^e_{010} + W^e_l\Gamma^e_{002}$ $+u^e_{A00l} + r^e_{Aojl} + \varepsilon^e_{Aijl}$, $C_l = \gamma^c_{000} + \gamma^c_{A001}T_l + W^c_l\Gamma^c_{002} + u^c_{Al}$	Standardized Noncentrality Parameter (λ): $\sqrt{\frac{\delta^2 P(1-P)nJL}{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw'^{ec}_2 r_2 + w'^{ec}_1 r_1)}}$ MDES: $\frac{M_{L-2-g}}{\sqrt{P(1-P)nJL}} \times$ $\sqrt{\kappa^2[(nJw^e_3 - w^e_1)\rho^e_3 + (nw^e_2 - w^e_1)\rho^e_2] + \psi_c[(nJw^c_3 - w^c_1)\rho^c_3 + (nw^c_2 - w^c_1)\rho^c_2] + (\kappa^2 w^e_1 + \psi_c w^c_1) - 2\kappa\sqrt{\psi_c}(nJw^{ec}_3 r_3 + nw'^{ec}_2 r_2 + w'^{ec}_1 r_1)}$	*L-2-g*

Table 2. Power and MDES estimates for specific scenarios with various assumptions: Two-level CRCETs and CRTs

Model		CRCETs						CRTs	
		Availability of cost information	No Cost variation	Cost varies among students and schools				Incremental cost	
				No Correlation	Positive	Negative	Mixed	Positive	Negative
Power	Unconditional	Level 2	0.485	0.441	0.559	0.365	0.553	0.803	0.290
		Levels 1 and 2	0.485	0.441	0.559	0.365	0.553	0.803	0.290
	Conditional	Level 2	0.776	0.723	0.844	0.623	0.837	0.978	0.511
		Levels 1 and 2	0.776	0.726	0.846	0.626	0.841	0.978	0.511
MDES	Unconditional	Level 2	0.729	0.773	0.665	0.868	0.669	0.399	0.399
		Levels 1 and 2	0.729	0.773	0.665	0.868	0.669	0.399	0.399
	Conditional	Level 2	0.516	0.549	0.471	0.617	0.476	0.282	0.282
		Levels 1 and 2	0.516	0.547	0.470	0.614	0.473	0.282	0.282

Note. (1) To compute power, we assume $\Delta E = 0.4$ and $\Delta C = 0.3$ when the incremental cost is positive and $\Delta E = 0.2$ and $\Delta C = -0.1$ when the incremental cost is negative. Therefore, under the assumption κ = *2,* $\delta_{CRCET} = 0.5$ in both scenarios. (2) To compute MDES, we assume *power = 0.8.* (3) Under the assumptions: *n = 60, J = 50,* $\rho_e = \rho_c = 0.23$*,* $w_1^e = w_1^c = w_2^e = w_2^c = w_1^{ec} = w_2^{ec} = 0.5$*,* $w'^{ec}_1 = 0.8$, $P = 0.5$, q = 1, and a two-sided test with α = 0.05. (4) No cost variation indicates $\psi_c = 0$; When cost varies among students and schools, we assume the total variance of cost is half as large as the total variance of effectiveness measures (i.e., $\psi_e = 2\psi_c$); No correlation indicates $r_1 = r_2 = 0$; Positive indicates $r_1 = r_2 = 0.1$; Negative indicates $r_1 = r_2 = -0.1$; Mixed indicates $r_1 = -0.1$ and $r_2 = 0.1$. (5) All the power and MDES for CRCETs were computed using PowerUp!-CEA. The Online Appendix Tables C1 and C2 illustrate the parameters used in the Two-level CRCET software package to generate the power and MDES estimates for conditional models when both level-1 and level-2 cost data are available, and the covariance between cost and effectiveness measures are positive, respectively. (6) The Online Appendix Table C5 illustrates the parameters used in the Two-level CRCET software package to generate the power for the conditional model when only level-2 cost data are available, and the correlation is positive. (7) The power and MDES for CRTs were computed using PowerUpR (Bulus et al., 2021).

Table 3. Power and MDES estimates for specific scenario with various assumptions: Three-level CRCETs and CRTs

	Model	CRCETs					CRTs	
		Availability of cost information	No Cost variation	Cost varies among students and schools			Incremental Cost	
				No Correlation	Positive	Negative	Mixed	Positive	Negative
Power	Unconditional	Level 3	0.556	0.508	0.587	0.421	0.581	0.919	0.390
		Levels 2 and 3	0.556	0.508	0.587	0.421	0.581	0.919	0.390
		Levels 1, 2, and 3	0.556	0.508	0.587	0.421	0.581	0.919	0.390
	Conditional	Level 3	0.844	0.788	0.848	0.681	0.840	0.997	0.661
		Levels 2 and 3	0.844	0.796	0.856	0.688	0.848	0.997	0.661
		Levels 1, 2, and 3	0.844	0.800	0.869	0.701	0.864	0.997	0.661
MDES	Unconditional	Level 3	0.667	0.707	0.643	0.796	0.647	0.334	0.334
		Levels 2 and 3	0.667	0.707	0.643	0.796	0.647	0.334	0.334
		Levels 1, 2, and 3	0.667	0.707	0.643	0.796	0.647	0.334	0.334
	Conditional	Level 3	0.472	0.508	0.469	0.576	0.474	0.236	0.236
		Levels 2 and 3	0.472	0.502	0.463	0.572	0.469	0.236	0.236
		Levels 1, 2, and 3	0.472	0.500	0.455	0.563	0.458	0.236	0.236

Note. (1) To compute power, we assume $\Delta E = 0.4$ and $\Delta C = 0.3$ when the incremental cost is positive and $\Delta E = 0.2$ and $\Delta C = -0.1$ when the incremental cost is negative. Therefore, under the assumption κ = *2,* $\delta_{CRCET} = 0.5$ in both scenarios. (2) To compute MDES, we assume *power = 0.8.* (3) Under the assumptions: *n =25, J = 2, L=60,* $\rho_2^e = \rho_2^c = 0.08$, $\rho_3^e = \rho_3^c = 0.15$, $w_1^e = w_1^c = w_2^e = w_2^c = w_3^e = w_3^c = w_1^{ec} = w_2^{ec} = w_3^{ec} = 0.5$, $w'^{ec}_1 = w_2'^{ec} = 0.8$, $P = 0.5$, $q = 1$, and a two-sided test with $\alpha = 0.05$. (4) No cost variation indicates $\psi_c = 0$; When cost varies among students and schools, we assume the total variance of cost is half as large as the total variance of effectiveness measures (i.e., $\psi_e = 2\psi_c$); No correlation indicates $r_1 = r_2 = r_3 = 0$; Positive indicates $r_1 = 0.1$ and $r_3 = 0.07$; Negative indicates $r_1 = -0.1$ and $r_3 = -0.07$; Mixed indicates $r_1 = -0.1$ and $r_3 = 0.07$. We always assume $r_2 = -0.03$ for the Positive, Negative, and Mixed scenarios. (5) All the power and MDES for CRCETs were computed using PowerUp!-CEA. The Online Appendix Tables C3 and C4 illustrate the parameters used in the three-level CRCET software package to generate the power and MDES estimate for conditional models when cost data at all levels are available, and the correlations are mixed, respectively. (6) The power and MDES for CRTs were computed using PowerUpR (Bulus et al., 2021).

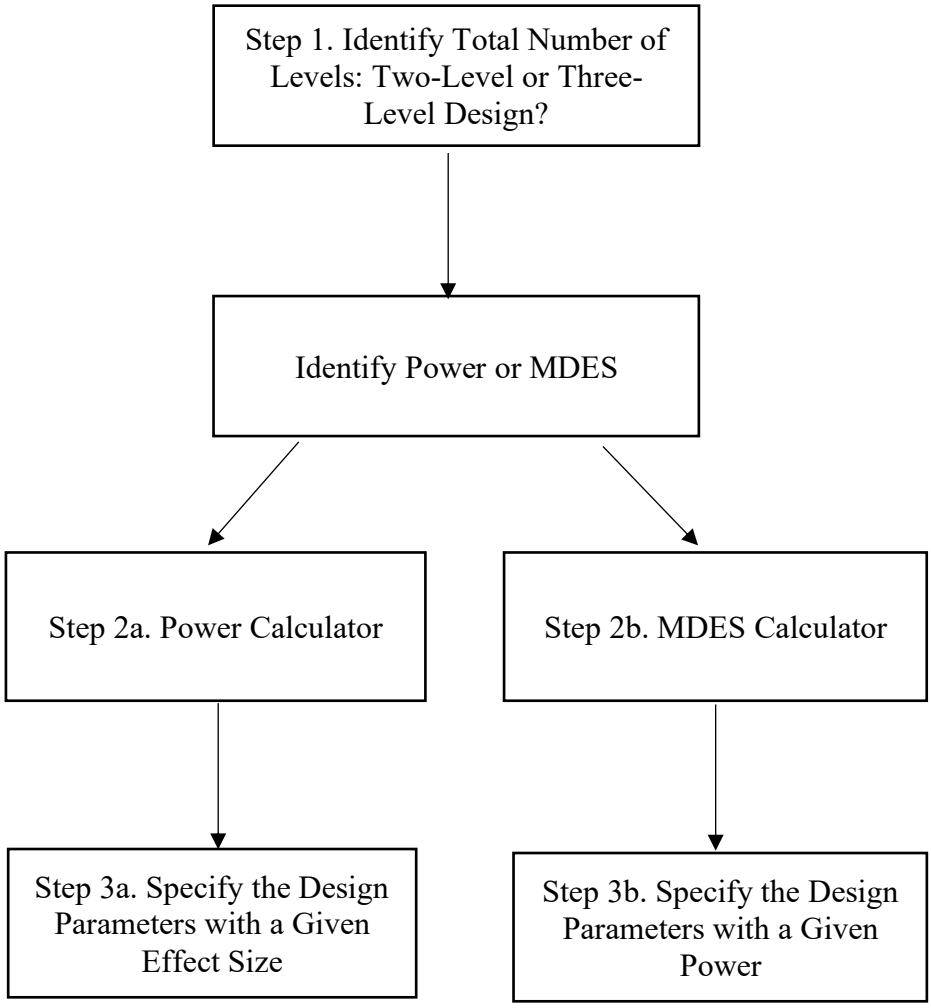

Figure 1. A three-step process to computer power or MDES for CRCETs using PowerUp!-CEA

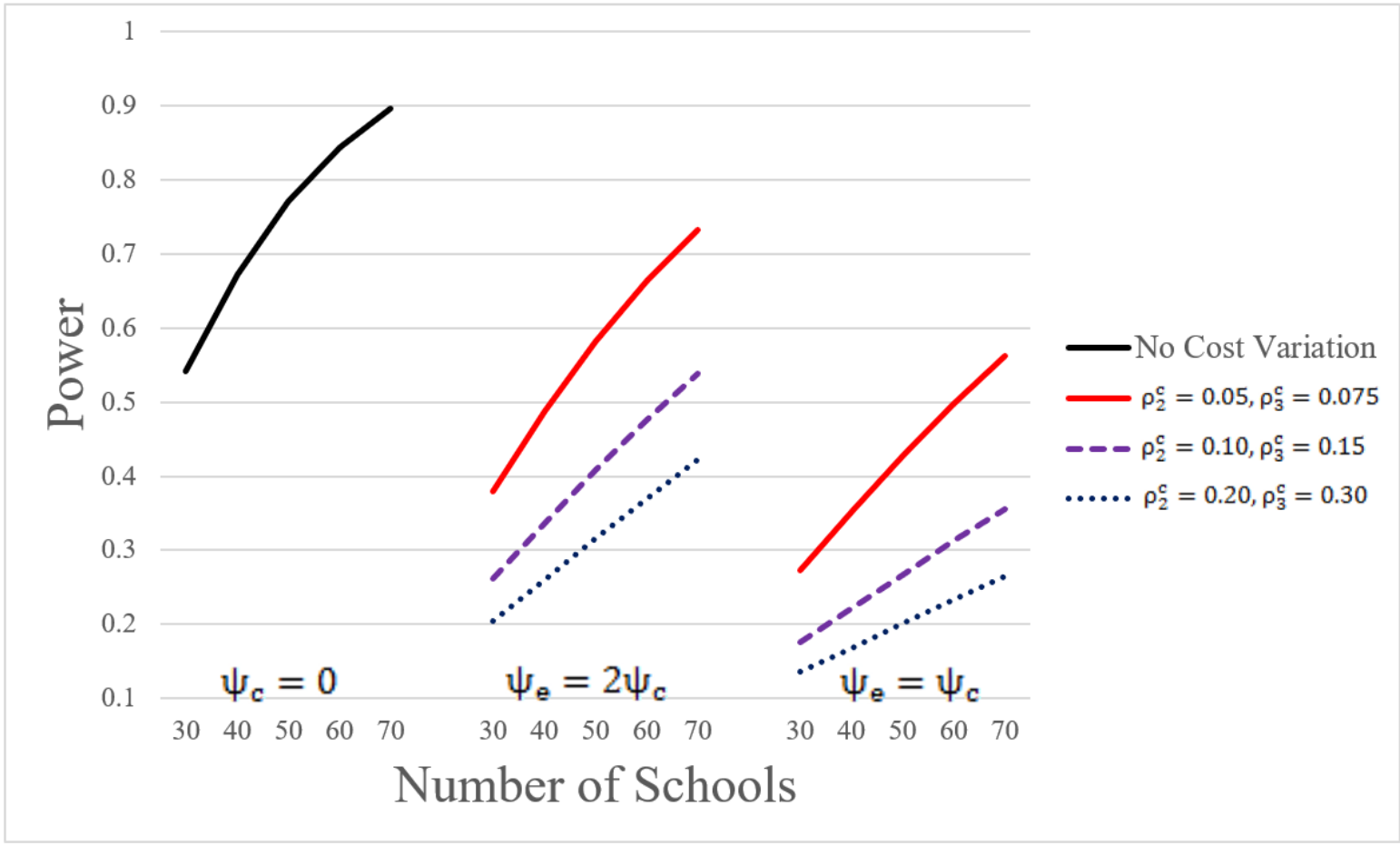

Figure 2. Effects of Sample Size and Cost Variation on Power

Note. Under the assumptions: $\kappa = 2$, $\delta = 0.5$, $J = 2$, $n = 25$, $\rho_2^e = 0.08$, $\rho_3^e = 0.15$, $r_1 = r_2 = r_2 = 0$, $P = 0.5$, no covariates, and a two-sided test with $\alpha = 0.05$.

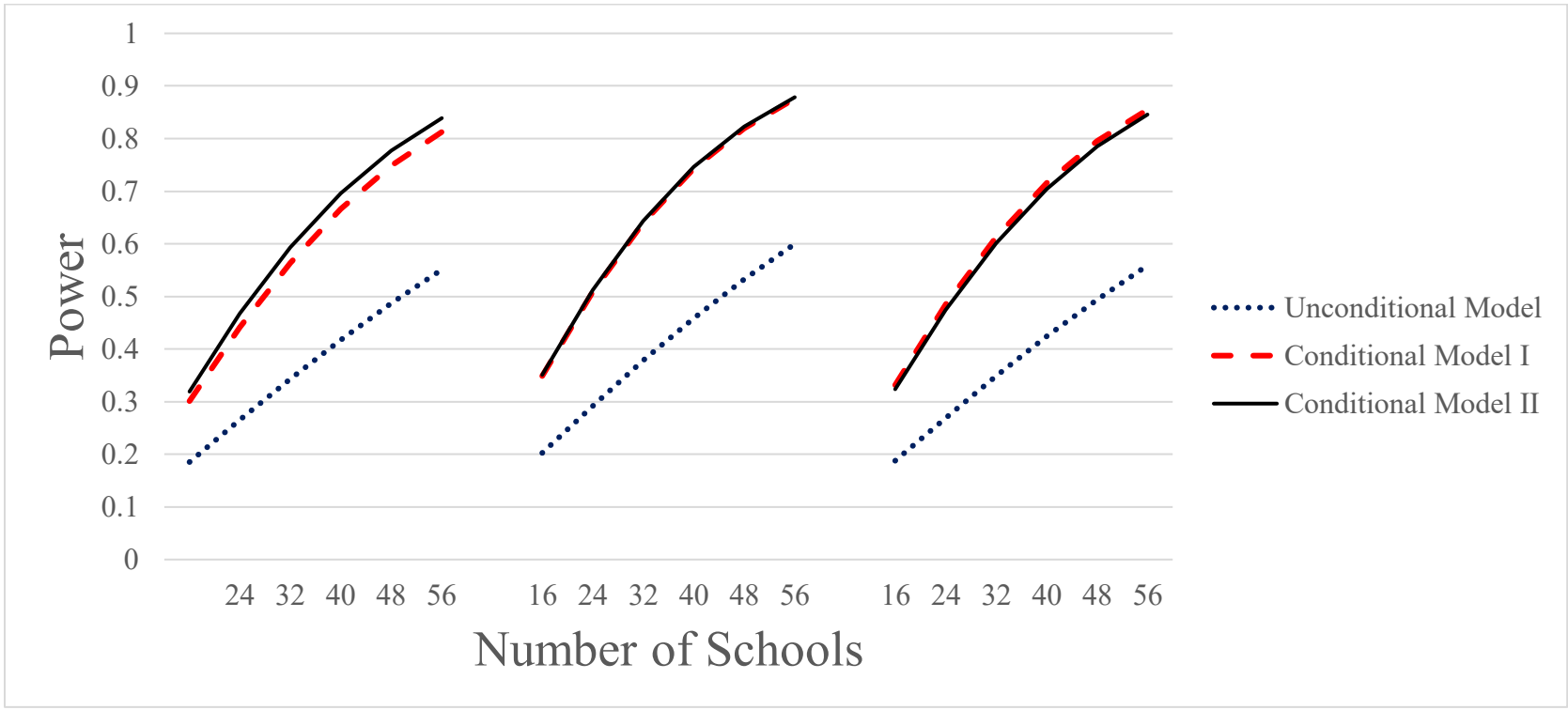

Figure 3. Effects of Covariate Adjustments on Power

Note. (1) The conditional model I assumes only school-level cost data are available; Conditional model II assumes cost data at three levels are all available. (2). Under the assumptions: $\kappa = 2,\ \delta = 0.5,\ J = 2,\ n = 25,\ \psi_e = \psi_c,\ \rho_2^e = \rho_2^c = 0.08,\ \rho_3^e = \rho_3^c = 0.15,\ w_1^e = w_2^e = w_3^e = 0.5, w_1^c = w_2^c = w_3^c = 0.5,\ q = 1,\ P = 0.5$, and a two-sided test with $\alpha = 0.05$. (3) For the three lines on the left-hand panel, we assume $r_1 = -0.1, r_2 = -0.03$, and $r_3 = 0.07$; for the three lines in the middle, we assume $r_1 = -0.1, r_2 = 0.03$, and $r_3 = 0.07$; and for the three lines on the right-hand panel, we assume $r_1 = -0.01, r_2 = 0.06$, and $r_3 = 0.07$.

Online Appendix A: Derivation of the Variance Formulas in Equations 11 and 20

When only level-2 cost information is available, we can estimate ΔC through a single-level regression, namely

$$C_j = \gamma_{00}^c + T_j \Delta C + r_j^c, \quad \text{(A1)}$$

where C_j is the cost for cluster j, γ_{00}^c is the grand mean of cost for the control group, and r_j^c is the error term for cost data. To compare the power for designs with or without level-1 cost information available, we rewrite r_j^c as a combination of error terms at the first and second levels using the same notations as those used in equation (3), namely

$$r_j^c = r_{0j}^c + \frac{\sum_{i=1}^{n} \varepsilon_{ij}^c}{n}, \quad \text{(A2)}$$

where ε_{ij}^c is the level-1 error term, and r_{0j}^c is the level-2 random effect. And thus, the variance of ΔC is

$$Var(\Delta C) = Var(\hat{\gamma}_{01}^c) = \frac{1}{P(1-P)nJ}(n\tau_c^2 + \sigma_c^2). \quad \text{(A3)}$$

Please note that when $\psi_c = 0$, indicating $\tau_c^2 = \sigma_c^2 = 0$, $Var(\Delta C)$ is equal to zero. Assume level-1 effectiveness data are available, we can still estimate ΔE through equation (3), and the variance of ΔE is (Raudenbush,1997)

$$Var(\Delta E) = Var(\hat{\gamma}_{01}^e) = \frac{1}{P(1-P)nJ}(n\tau_e^2 + \sigma_e^2). \quad \text{(A4)}$$

Similarly, the covariance between ΔE and ΔC is:

$$Cov(\Delta E, \Delta C) = \frac{1}{P(1-P)nJ}(n\tau_{ec} + \sigma_{ec}). \quad \text{(A5)}$$

Based on equation (1) we have the variance of $\widehat{INMB}$ (equation 12) as

$$Var\left(\widehat{INMB}\right) = \frac{1}{P(1-P)nJ}[\kappa^2(n\tau_e^2 + \sigma_e^2) + (n\tau_c^2 + \sigma_c^2) - 2\kappa(n\tau_{ec} + \sigma_{ec})]. \quad \text{(A6)}$$

When only level-2 cost data are available, and there are covariates incorporated in the analysis, we could estimate ΔC through a single-level regression with level-2 covariates, namely

$$C_j = \gamma_{00}^c + T_j \Delta C + Z_j^c \Gamma_{02}^c + r_{Aj}^c, \quad \text{(A7)}$$

where Z_j^c is a row vector of level-2 unit characteristics, Γ_{02}^c is a column vector of coefficients of level-2 unit characteristics, and r_{Aj}^c is the residual term. Subscript A indicates adjustment because of covariates. Again, we could rewrite r_{Aj}^c as combinations of error terms at the first and second levels, namely

$$r_{Aj}^c = r_{A0j}^c + \frac{\sum_{i=1}^n \varepsilon_{ij}^c}{n}. \quad \text{(A8)}$$

Then the variance of ΔC becomes

$$Var(\Delta C) = Var(\hat{\gamma}_{A01}^c) = \frac{1}{p(1-p)nJ}(n\tau_{Rc}^2 + \sigma_c^2). \quad \text{(A9)}$$

Assume level-1 effectiveness data are available, and thus we could use equation (14) to estimate the variance of ΔE as

$$Var(\Delta E) = Var(\hat{\gamma}_{A01}^e) = \frac{1}{p(1-p)nJ}(n\tau_{Re}^2 + \sigma_{Re}^2). \quad \text{(A10)}$$

Similarly, the covariance between ΔE and ΔC becomes

$$Cov(\Delta E,\ \Delta C) = \frac{1}{P(1-P)nJ}(n\tau_{Rec} + \sigma'_{Rec}). \quad \text{(A11)}$$

where σ'_{Rec} represents the covariance between cost and effectiveness data at level-1 that only considers the covariate effects for effectiveness measures. Therefore, we have the variance of $\widehat{INMB}$ (equation 21) as

$$Var\left(\widehat{INMB}\right) = \frac{1}{P(1-P)nJ}[\kappa^2(n\tau_{Re}^2 + \sigma_{Re}^2) + (n\tau_{Rc}^2 + \sigma_c^2) - 2\kappa(n\tau_{Rec} + \sigma'_{Rec})]. \quad \text{(A12)}$$

Online Appendix B: Three-Level Designs

Three-level models: Unconditional Model

Consider a three-level cluster design (e.g., students nested within classes, and classes nested within schools), where level-3 units (e.g., schools) are randomly assigned to treatment or control conditions and the treatment is at the third level. When the level-1 effectiveness data and cost data are available, three-level HLMs could be used to estimate the incremental effect and the incremental cost (i.e., ΔE and ΔC), namely

$$E_{ijl} = \gamma^e_{000} + T_l \Delta E + u^e_{00l} + r^e_{0jl} + \varepsilon^e_{ijl}, \quad \text{(B1)}$$

$$C_{ijl} = \gamma^c_{000} + T_l \Delta C + u^c_{00l} + r^c_{0jl} + \varepsilon^c_{ijl}, \quad \text{(B2)}$$

where E_{ijl} represents the effectiveness measure (e.g., achievement) for level-1 unit *i* in level-2 unit *j* within level-3 unit *l*; C_{ijl} represents the cost for level-1 unit *i* in level-2 unit *j* within level-3 unit *l*; T_l is a binary treatment indicator variable; and the level-1 error and random effects at level-2 and level-3 follow bivariate normal distributions, namely

$$\begin{pmatrix} u^e_{00l} \\ u^c_{00l} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega^2_e & \omega_{ec} \\ \omega_{ec} & \omega^2_c \end{pmatrix} \right), \begin{pmatrix} r^e_{0jl} \\ r^c_{0jl} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_e & \tau_{ec} \\ \tau_{ec} & \tau^2_c \end{pmatrix} \right), \text{ and } \begin{pmatrix} \varepsilon^e_{ijl} \\ \varepsilon^c_{ijl} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_e & \sigma_{ec} \\ \sigma_{ec} & \sigma^2_c \end{pmatrix} \right). \quad \text{(B3)}$$

Again, let NMB_{ijl} represent the NMB for level-1 unit *i* in level-2 unit *j* within level-3 unit *l*, we can reconstruct equations (B1) and (B2) as

$$NMB_{ijl} = \kappa E_{ijl} - C_{ijl} = \pi_{000} + \pi_{001} T_l + u_{00l} + r_{0jl} + \varepsilon_{ijl}, \quad \text{(B4)}$$

where $\pi_{000} = \kappa\gamma^e_{000} - \gamma^c_{000}$, $\pi_{001} = \kappa\Delta E - \Delta C$, $u_{00l} = \kappa u^e_{00l} - u^c_{00l}$, $r_{0jl} = \kappa r^e_{0jl} - r^c_{0jl}$, $\varepsilon_{ijl} = \kappa\varepsilon^e_{ijl} - \varepsilon^c_{ijl}$, $u_{00l} \sim N(0, \omega^2)$, $r_{0jl} \sim N(0, \tau^2)$, and $\varepsilon_{ijl} \sim N(0, \sigma^2)$. The parameter of interest now is π_{001}, representing the INMB of the treatment. When $\pi_{001} > 0$, it indicates the treatment is cost-effective, when $\pi_{001} < 0$, it indicates the treatment is not cost-effective.

Suppose there are L level-3 units, J level-2 units within each level-3 unit, and n level-1 units within each level-2 unit. The total number of level-1 units is nJL. Also, suppose there are L_T level-3 units in the treatment group and L_C in the control condition. Define $P = \frac{L_T}{L}$, is the proportion of level-3 units in the treatment group, then the variance of $\hat{\pi}_{001}$ is (Konstantopoulos, 2008a)

$$Var(\hat{\pi}_{001}) = \frac{1}{P(1-P)nJL}(nJ\omega^2 + n\tau^2 + \sigma^2), \quad \text{(B5)}$$

where $\omega^2 = \kappa^2\omega_e^2 - 2\kappa\omega_{ec} + \omega_c^2$, $\tau^2 = \kappa^2\tau_e^2 - 2\kappa\tau_{ec} + \tau_c^2$, and $\sigma^2 = \kappa^2\sigma_e^2 - 2\kappa\sigma_{ec} + \sigma_c^2$. The non-centrality parameter is

$$\lambda = \pi_{001}\sqrt{\frac{P(1-P)nJL}{(nJ\omega^2+n\tau^2+\sigma^2)}}. \quad \text{(B6)}$$

Similarly, let $\psi_e = \sigma_e^2 + \tau_e^2 + \omega_e^2$ and $\psi_c = \sigma_e^2 + \tau_e^2 + \omega_e^2$ represent the total variance of effective and cost measures. If we assume the effectiveness measure is standardized with a mean of zero and a standard deviation of one (i.e., $\psi_e = 1$) and then define the effect size as $\delta = INMB$, the standardized non-centrality parameter is

$$\lambda = \delta\sqrt{\frac{P(1-P)nJL}{\kappa^2[(nJ-1)\rho_3^e+(n-1)\rho_2^e]+\psi_c[(nJ-1)\rho_3^c+(n-1)\rho_2^c]+(\kappa^2+\psi_c)-2\kappa\sqrt{\psi_c}(nJr_3+nr_2+r_1)}}, \quad \text{(B7)}$$

where $\rho_3^e = \frac{\omega_e^2}{\psi_e}$ and $\rho_2^e = \frac{\tau_e^2}{\psi_e}$ are the ICCs of effectiveness data at the third and second levels, respectively; $\rho_3^c = \frac{\omega_c^2}{\psi_c}$ and $\rho_2^c = \frac{\tau_c^2}{\psi_c}$ are ICCs of cost data at the third and second levels, respectively; and $r_1 = \frac{\sigma_{ec}}{\sqrt{\psi_e\psi_c}}$ and $r_2 = \frac{\tau_{ec}}{\sqrt{\psi_e\psi_c}}$, and $r_3 = \frac{\omega_{ec}}{\sqrt{\psi_e\psi_c}}$ are the standardized covariance between cost and effectiveness at the first, second, and third levels, respectively. Then, power is defined as:

Power = 1 – H [*c*(α */2, L-2*), (*L-2*), *λ*] + H [*-c*(α */2, L-2*), (*L-2*), *λ*]. (B8)

When only level-3 cost data are available, but level-1 effectiveness data are still available, we could estimate ΔC through a single-level regression, namely

$$C_l = \gamma^c_{000} + T_l \Delta C + u^c_l, \tag{B9}$$

where C_l is the cost for cluster l, γ^c_{000} is the grand mean of cost for the control group, and u^c_l is the error term. To compare the power for designs with or without level-1 cost information, we could rewrite u^c_l as combinations of error terms at the first, second, and third levels using the same notations as those used in equation (B2), namely

$$u^c_l = u^c_{00l} + \frac{\sum_{j=1}^{J} r^c_{0jl}}{J} + \frac{\sum_{j=1}^{J}\sum_{i=1}^{n} \varepsilon^c_{ijl}}{nJ}, \tag{B10}$$

where ε^c_{ijl} is the level-1error term, r^c_{0jl} is the level-2 random effect, and u^c_{00l} is the level-3 random effect. And thus, we could write the variance of ΔC as

$$Var(\Delta C) = Var(\hat{\gamma}^c_{001}) = \frac{1}{P(1-P)nJL}(nJ\omega_c^2 + n\tau_c^2 + \sigma_c^2). \tag{B11}$$

Assuming level-1 effectiveness data are available, we could use equation (B3) to estimate ΔE. Specifically, according to Konstantopoulos (2008), the variance of ΔE is

$$Var(\Delta E) = Var(\hat{\gamma}^e_{001}) = \frac{1}{P(1-P)nJL}(nJ\omega_e^2 + n\tau_e^2 + \sigma_e^2). \tag{B12}$$

And then the covariance between ΔE and ΔC is

$$Cov(\Delta E, \Delta C) = \frac{1}{P(1-P)nJL}(nJ\omega_{ec} + n\tau_{ec} + \sigma_{ec}). \tag{B13}$$

Based on equation (1) in the main text, we can get the variance of $\widehat{INMB}$ as

$$Var\left(\widehat{INMB}\right) = \frac{1}{P(1-P)nJL}[\kappa^2(nJ\omega_e^2 + n\tau_e^2 + \sigma_e^2) + (nJ\omega_c^2 + n\tau_c^2 + \sigma_c^2) - 2\kappa(nJ\omega_{ec} + n\tau_{ec} + \sigma_{ec})] \tag{B14}$$

Again, define the standardized effect size as $\delta = INMB$ and assume the effectiveness measures are standardized with means of zero and standard deviations of one (i.e., $\psi_e = 1$), we have the standardize the non-centrality parameter as

$$\lambda = \delta\sqrt{\frac{P(1-P)nJL}{\kappa^2[(nJ-1)\rho_3^e+(n-1)\rho_2^e]+\psi_c[(nJ-1)\rho_3^c+(n-1)\rho_2^c]+(\kappa^2+\psi_c)-2\kappa\sqrt{\psi_c}(nJr_3+nr_2+r_1)}}. \quad \text{(B15)}$$

Note that equation (B15) is identical to equation (B7). Similarly, when level-1 cost data are not available, but level-2 cost data are available, we could estimate the INMB through a two-level (e.g., classes nested within schools) model, where the new level-1 error term is a combination of level-1 (e.g., students) and level-2 (e.g., classes) errors. Then, the non-centrality is also identical to equations (B7) and (B15), indicating the power of detecting the cost-effectiveness of treatment is the same for unconditional models, regardless of whether level-1 or level-2 cost data are available or not for three-level CRCETs.

Three-Level Cluster Design: Covariate Effects

When there are covariates incorporated in the analysis and level-1 cost data are available, we can still use the three-level HLMs to estimate the incremental effect and the incremental cost of an intervention, namely

$$e_{ijl} = \gamma_{000}^e + T_l\Delta E + X_{ijl}^e\Gamma_{100}^e + Z_{jl}^e\Gamma_{010}^e + W_l^e\Gamma_{002}^e + u_{A00l}^e + r_{A0jl}^e + \varepsilon_{Aijl}^e, \quad \text{(B16)}$$

$$c_{ijl} = \gamma_{000}^c + T_l\Delta C + X_{ijl}^c\Gamma_{100}^c + Z_{jl}^c\Gamma_{010}^c + W_l^c\Gamma_{002}^c + u_{A00l}^c + r_{A0jl}^c + \varepsilon_{Aijl}^c, \quad \text{(B17)}$$

where X_{ijl}^e and X_{ijl}^c are row vectors of level-1 unit characteristics, and Γ_{100}^e and Γ_{100}^c are column vectors of coefficients of level-1 unit characteristics; Z_{jl}^e and Z_{jl}^c row vectors of level-2 unit characteristics, and Γ_{010}^e and Γ_{010}^c are column vectors of coefficients of level-2 unit characteristics; W_l^e and W_l^c are row vectors of level-3 unit characteristics, and Γ_{002}^e and Γ_{002}^c are column vectors of coefficients of level-3 unit characteristics. Subscript *A* indicates adjustment

because of covariates. The level-1 error and random effects at level-2 and level-3 follow bivariate normal distributions

$$\begin{pmatrix} u^{e}_{A00k} \\ u^{c}_{A00k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega^{2}_{Re} & \omega_{Rec} \\ \omega_{Rec} & \omega^{2}_{Rc} \end{pmatrix} \right), \begin{pmatrix} r^{e}_{A0jk} \\ r^{c}_{A0jk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^{2}_{Re} & \tau_{Rec} \\ \tau_{Rec} & \tau^{2}_{Rc} \end{pmatrix} \right), \text{ and}$$

$$\begin{pmatrix} \varepsilon^{e}_{Aijk} \\ \varepsilon^{c}_{Aijk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^{2}_{Re} & \sigma_{Rec} \\ \sigma_{Rec} & \sigma^{2}_{Rc} \end{pmatrix} \right),$$

(B18)

where subscript *R* indicates residual variance or residual covariance. Then, the NMB for level-1 unit *i* in level-2 unit *j* in cluster *l* becomes

$$NMB_{ijl} = \pi_{000} + \pi_{A001}T_{l} + X_{ijl}\Gamma_{100} + Z_{jl}\Gamma_{010} + W_{l}\Gamma_{002} + u_{A00l} + r_{Aojl} + \varepsilon_{Aijl}. \quad \text{(B19)}$$

And the non-centrality parameter becomes

$$\lambda = \delta \times \sqrt{\frac{P(1-P)nJL}{\kappa^{2}\left[(nJw^{e}_{3}-w^{e}_{1})\rho^{e}_{3}+(nw^{e}_{2}-w^{e}_{1})\rho^{e}_{2}\right]+\psi_{c}\left[(nJw^{c}_{3}-w^{c}_{1})\rho^{c}_{3}+(nw^{c}_{2}-w^{c}_{1})\rho^{c}_{2}\right]+(\kappa^{2}w^{e}_{1}+\psi_{c}w^{c}_{1})-2\kappa\sqrt{\psi_{c}}(nJw^{ec}_{3}r_{3}+nw^{ec}_{2}r_{2}+w^{ec}_{1}r_{1})}},$$

(B20)

where w^{e}_{1}, w^{e}_{2}, w^{e}_{3} represent the unexplained variance of effectiveness at the first, second, and third levels, respectively; w^{c}_{1}, w^{c}_{2}, and w^{c}_{3} represent the unexplained variance of cost at the first, second, and third levels, respectively; w^{ec}_{1}, w^{ec}_{2}, and w^{ec}_{3} represent the unexplained covariance between cost and effectiveness at the first, second, and third levels, respectively. Specifically, $w^{e}_{1} = \frac{\sigma^{2}_{Re}}{\sigma^{2}_{e}}$, $w^{e}_{2} = \frac{\tau^{2}_{RTe}}{\tau^{2}_{Te}}$, $w^{e}_{3} = \frac{\omega^{2}_{Re}}{\omega^{2}_{e}}$, $w^{c}_{1} = \frac{\sigma^{2}_{Rc}}{\sigma^{2}_{c}}$, $w^{c}_{2} = \frac{\tau^{2}_{RTc}}{\tau^{2}_{Tc}}$, $w^{c}_{3} = \frac{\omega^{2}_{Rc}}{\omega^{2}_{c}}$, $w^{ec}_{1} = \frac{\sigma_{Rec}}{\sigma_{ec}}$, $w^{ec}_{2} = \frac{\tau_{RTec}}{\tau_{Tec}}$, and $w^{ec}_{3} = \frac{\omega_{Rec}}{\omega_{ec}}$. Note that we assume group-mean centering of level-1 and level-2 covariates so that they could only explain a proportion of the variance or covariance corresponding levels. Then, power is defined as

$$Power = 1 - \mathrm{H}\left[c(\alpha/2,\ L\text{-}2\text{-}g),\ (L\text{-}2\text{-}g),\ \lambda_{A}\right] + \mathrm{H}\left[-c(\alpha/2,\ L\text{-}2\text{-}g),\ (L\text{-}2\text{-}g),\ \lambda_{A}\right], \quad \text{(B21)}$$

where g is the number of covariates at the third level. All the other terms have been defined previously. The minimum detectable effect size (MDES) is

$$MDES(\delta) = \frac{M_v}{\sqrt{P(1-P)nJL}} \times \sqrt{\kappa^2[(nJw_3^e - w_1^e)\rho_3^e + (nw_2^e - w_1^e)\rho_2^e] + \psi_c[(nJw_3^c - w_1^c)\rho_3^c + (nw_2^c - w_1^c)\rho_2^c] + (\kappa^2 w_1^e + \psi_c w_1^c) - 2\kappa\sqrt{\psi_c}(nJw_3^{ec} r_3 + nw_2^{ec} r_2 + w_1^{ec} r_1)}. \quad \text{(B22)}$$

When only level-3 cost data are available, and covariates are incorporated in the analysis, we could estimate ΔC through a single-level regression with level-3 covariates, namely

$$C_l = \gamma_{000}^c + T_l \Delta C + W_l^c \Gamma_{002}^c + u_{Al}^c, \quad \text{(B23)}$$

where W_l^c is a row vector of level-3 unit characteristics, Γ_{100}^c is a column vector of coefficients of level-3 unit characteristics, and u_{Al}^c is the third level error term for cost data. Subscript *A* indicates adjustment because of covariates. Again, we could rewrite u_{Al}^c as combinations of error terms at the first, second, and third levels, namely

$$u_{Al}^c = u_{A00l}^c + \frac{\sum_{j=1}^J r_{0jl}^c}{J} + \frac{\sum_{j=1}^J \sum_{i=1}^n \varepsilon_{ijl}^c}{nJ}. \quad \text{(B24)}$$

Then the variance of ΔC becomes

$$Var(\Delta C) = Var(\hat{\gamma}_{A001}^c) = \frac{1}{P(1-P)nJL}(nJ\omega_{Rc}^2 + n\tau_c^2 + \sigma_c^2). \quad \text{(B25)}$$

Assume individual level effectiveness measure is available, and thus we could use equation (B16) to estimate the variance of ΔE as

$$Var(\Delta E) = Var(\hat{\gamma}_{A001}^e) = \frac{1}{pP(1-P)nJL}(nJ\omega_{Re}^2 + n\tau_{Re}^2 + \sigma_{Re}^2). \quad \text{(B26)}$$

And then the covariance between ΔE and ΔC is

$$Cov(\Delta E,\ \Delta C) = \frac{1}{P(1-P)nJL}(nJ\omega_{Rec} + n\tau'_{Rec} + \sigma'_{Rec}), \quad \text{(B27)}$$

where τ'_{Rec} and σ'_{Rec} represent the covariance between cost and effectiveness measures at level-2 and level-1 that only consider the covariate effects for effectiveness measures. Therefore, we have the variance of INMB as

$$Var\left(\widehat{INMB}\right) = \frac{1}{P(1-P)nJL}[\kappa^2(nJ\omega_{Re}^2 + n\tau_{Re}^2 + \sigma_{Re}^2) + (nJ\omega_{Rc}^2 + n\tau_c^2 + \sigma_c^2) - 2\kappa(nJ\omega_{Rec} + n\tau'_{Rec} + \sigma'_{Rec})]. \quad \text{(B28)}$$

Again, assume the effectiveness measure is standardized with a mean of zero and a standard deviation of one (i.e., $\psi_e = 1$) and define the standardized effect size as $\delta = INMB$, then the standardized the non-centrality parameter becomes

$$\lambda = \delta\sqrt{\frac{P(1-P)nJL}{\kappa^2[(nJw_3^e - w_1^e)\rho_3^e + (nw_2^e - w_1^e)\rho_2^e] + \psi_c[(nJw_3^c - 1)\rho_3^c + (n-1)\rho_2^c] + (\kappa^2 w_1^e + \psi_c) - 2\kappa\sqrt{\psi_c}(nJw_3^{ec}r_3 + nw\prime_2^{ec}r_2 + w\prime_1^{ec}r_1)}}, \quad \text{(B29)}$$

which is not identical to equation (B21), because w'^{ec}_1 and w'^{ec}_2 only take account of the effects of covariate adjustments in the effectiveness model.

Similarly, when level-1 cost data are not available but level-2 cost data available, we could use a two-level HLM estimate INMB. Assume level-2 covariates could explain a proportion variation of the outcome variance at level-2, the non-centrality parameter becomes

$$\lambda = \delta\sqrt{\frac{P(1-P)nJL}{\kappa^2[(nJw_3^e - w_1^e)\rho_3^e + (nw_2^e - w_1^e)\rho_2^e] + \psi_c[(nJw_3^c - 1)\rho_3^c + (nw_2^c - 1)\rho_2^c] + (\kappa^2 w_1^e + \psi_c) - 2\kappa\sqrt{\psi_c}(nJw_3^{ec}r_3 + nw_2^{ec}r_2 + w'^{ec}_1 r_1)}}, \quad \text{(B30)}$$

where w'^{ec}_1 only takes account of the effects of covariate adjustments in the effectiveness model.

Online Appendix C: Demonstrations using PowerUp!-CEA

Table C1. Demonstration of Power Computation for Two-level CRCETs: Level-1 Cost Data are Available

Model 2.1: Power Calculator for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Effect Size Difference	0.50	INMB standardized by the standard deviation of effectiveness data
Willingness to Pay (κ)	2.00	
P	0.50	Proportion of Level 2 units randomized to treatment: J_T / J
ψ_c/ψ_e	0.50	Raito of the total variance of cost data to the total variance of effectiveness data
Parameters for Effectiveness Data		
ρ^e	0.23	Proportion of variance in effectiveness measures that is between clusters
R^2_{1e}	0.50	Proportion of variance of effectiveness data explained by level-1 covariates
R^2_{2e}	0.50	Proportion of variance of effectiveness data explained by level-2 covariates
Parameters for Cost Data		
ρ^c	0.23	Proportion of variance in cost measures that is between clusters
R^2_{1c}	0.50	Proportion of variance of cost data explained by level-1 covariates
R^2_{2c}	0.50	Proportion of variance of cost data explained by level-2 covariates
Parameters for Covariation between Effectiveness Data and Cost Data		
r_1	0.10	Standardized covariance between the effectiveness data and cost data at level-1
r_2	0.10	Standardized covariance between the effectiveness data and cost data at level-2
R^2_{1ec}	0.50	Proportion of the covariance explained by level-1 covariates
R^2_{2ec}	0.50	Proportion of the covariance explained by level-2 covariates
g*	1	Number of Level 2 covariates
n (Average Cluster Size)	50	Mean number of Level 1 units per Level 2 cluster (geometric mean recommended)
J (Sample Size [# of Clusters])	60	Number of Level 2 units
Noncentrality Parameter	**3.03**	Automatically computed from the above assumptions
Power (1-β)	**0.846**	Statistical power (1-probability of a Type II error)

Note: (1)The parameters in yellow cells need to be specified. The power will be calculated automatically. (2) We always assume the effectiveness data are standardized with mean zero and standard deviation one.

Table C2. Demonstration of MDES Computation for Two-level CRCETs: Level-1 Cost Data are Available

Model 2.1: MDES Calculator for for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Willingness to Pay (κ)	2.00	
P	0.50	Proportion of Level 2 units randomized to treatment: J_T / J
ψ_c/ψ_e	0.50	Raito of the total variance of cost data to the total variance of effectiveness data
Parameters for Effectiveness data		
ρ^e	0.23	Proportion of variance in effectivness measures that is between clusters
R^2_{1e}	0.50	Proportion of variance of effectiveness data explained by level-1 covariates
R^2_{2e}	0.50	Proportion of variance of effectiveness data explained by level-2 covariates
Parameters for Cost Data		
ρ^c	0.23	Proportion of variance in cost measures that is between clusters
R^2_{1c}	0.50	Proportion of variance of cost data explained by level-1 covariates
R^2_{2c}	0.50	Proportion of variance of cost data explained by level-2 covariates
Parameters for Covariation between Effectiveness data and Cost Data		
r_1	0.10	Standardized covariance between the effectiveness data and cost data at level-1
r_2	0.10	Standardized covariance between the effectiveness data and cost data at level-2
R^2_{1ec}	0.50	Proportion of the covariance explained by level-1 covariates
R^2_{2ec}	0.50	Proportion of the covariance explained by level-2 covariates
g*	1	Number of Level 2 covariates
n (Average Cluster Size)	50	Mean number of Level 1 units per Level 2 cluster (geometric mean recommended)
J (Sample Size [# of Clusters])	60	Number of Level 2 units
M (Multiplier)	2.85	Computed from T_1 and T_2
T_1 (Precision)	2.00	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.470**	Minimum Detectable Effect Size Standardized by the Standard Deviation of the Effectiveness Data

Note: (1)The parameters in yellow cells need to be specified. The MDES will be calculated automatically. (2) We always assume the effectiveness data are standardized with mean zero and standard deviation one.

Table C3. Demonstration of Power Computation for Three-level CRCETs: Level-1 and Level-2 Cost Data are Available

Model 2.2: Power Calculator for for Three-Level Cluster Random Assignment Design (CRA3_3)— Treatment at Level 3		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Effect Size Difference	0.50	INMB standardized by the standard deviation of effectiveness data
Willingness to Pay (κ)	2.00	
P	0.50	Proportion of Level 3 units randomized to treatment: L_T / L
ψ_c/ψ_e	0.50	Raito of the total variance of cost data to the total variance of effectiveness data
Parameters for Effectiveness Data		
ρ_2^e	0.08	Proportion of variance in effectiveness measures among Level 2 units
ρ_3^e	0.15	Proportion of variance in effectiveness measures among Level 3 Units
R_{1e}^2	0.50	Proportion of variance of effectiveness data explained by level-1 covariates
R_{2e}^2	0.50	Proportion of variance of effectiveness data explained by level-2 covariates
R_{3e}^2	0.50	Proportion of variance of effectiveness data explained by level-3 covariates
Parameters for Cost Data		
$\boldsymbol{\rho_2^c}$	0.08	Proportion of variance in cost measures among Level 2 units
ρ_3^c	0.15	Proportion of variance in cost measures among Level 3 Units
R_{1c}^2	0.50	Proportion of variance of cost data explained by level-1 covariates
R_{2c}^2	0.50	Proportion of variance of cost data explained by level-2 covariates
R_{3c}^2	0.50	Proportion of variance of cost data explained by level-3 covariates
Parameters for Covariation between Effectiveness Data and Cost Data		
r_1	-0.10	Standardized covariance between the effectiveness data and cost data at level-1
r_2	-0.03	Standardized covariance between the effectiveness data and cost data at level-2
r_3	0.07	Standardized covariance between the effectiveness data and cost data at level-3
R_{1ec}^2	0.50	Proportion of the covariance explained by level-1 covariates
R_{2ec}^2	0.50	Proportion of the covariance explained by level-2 covariates
R_{3ec}^2	0.50	Proportion of the covariance explained by level-3 covariates
g*	1	Number of Level 3 covariates
n (Average Sample Size for Level 1)	25	Mean number of Level 1 units per Level 2 unit (geometric mean recommended)
J (Average Sample Size for Level 2)	2	Mean number of Level 2 units per Level 3 unit (geometric mean recommended)
L (Sample Size [# of Level 3 units])	60	The number of Level 3 units in the sample
Noncentrality Parameter	**3.112**	Automatically computed from the above assumptions
Power (1-β)	**0.864**	Statistical power (1-probability of a Type II error)

Note: (1)The parameters in yellow cells need to be specified. The power will be calculated automatically. (2) We always assume the effectiveness data are standardized with mean zero and standard deviation one.

Table C4. Demonstration of MDES Computation for Three-level CRCETs: Level-1 and Level-2 Cost Data are Available

Model 2.2: MDES Calculator for for Three-Level Cluster Random Assignment Design (CRA3_3)— Treatment at Level 3		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Willingness to Pay (κ)	2.00	
P	0.50	Proportion of Level 3 units randomized to treatment: L_T / L
ψ_c/ψ_e	0.50	Raito of the total variance of cost data to the total variance of effectiveness data
Parameters for Effectiveness Data		
ρ_2^e	0.08	Proportion of variance in effectiveness measures among Level 2 units
ρ_3^e	0.15	Proportion of variance in effectiveness measures among Level 3 Units
R_{1e}^2	0.50	Proportion of variance of effectiveness data explained by level-1 covariates
R_{2e}^2	0.50	Proportion of variance of effectiveness data explained by level-2 covariates
R_{3e}^2	0.50	Proportion of variance of effectiveness data explained by level-3 covariates
Parameters for Cost Data		
$\boldsymbol{\rho_2^c}$	0.08	Proportion of variance in cost measures among Level 2 units
ρ_3^c	0.15	Proportion of variance in cost measures among Level 3 Units
R_{1c}^2	0.50	Proportion of variance of cost data explained by level-1 covariates
R_{2c}^2	0.50	Proportion of variance of cost data explained by level-2 covariates
R_{3c}^2	0.50	Proportion of variance of cost data explained by level-3 covariates
Parameters for Covariation between Effectiveness Data and Cost Data		
r_1	-0.10	Standardized covariance between the effectiveness data and cost data at level-1
r_2	-0.03	Standardized covariancen between the effectiveness data and cost data at level-2
r_3	0.07	Standardized covariance between the effectiveness data and cost data at level-3
R_{1ec}^2	0.50	Proportion of the covariance explained by level-1 covariates
R_{2ec}^2	0.50	Proportion of the covariance explained by level-2 covariates
R_{3ec}^2	0.50	Proportion of the covariance explained by level-3 covariates
g*	1	Number of Level 3 covariates
n (Average Sample Size for Level 1)	25	Mean number of Level 1 units per Level 2 unit (geometric mean recommended)
J (Average Sample Size for Level 2)	2	Mean number of Level 2 units per Level 3 unit (geometric mean recommended)
L (Sample Size [# of Level 3 units])	60	The number of Level 3 units in the sample
M (Multiplier)	2.85	Computed from T_1 and T_2
T_1 (Precision)	2.002	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.848	Determined from given power level
MDES	**0.458**	Minimum Detectable Effect Size Standardized by the Standard Deviation of the Effectiveness Data

Note: (1)The parameters in yellow cells need to be specified. The power will be calculated automatically. (2) We always assume the effectiveness data are standardized with mean zero and standard deviation one.

Table C5. Demonstration of Power Computation for Two-level CRCETs: Only Level-2 Cost Data are Available

Model 2.1: Power Calculator for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Effect Size Difference	0.50	INMB standardized by the standard deviation of effectiveness data
Willingness to Pay (κ)	2.00	
P	0.50	Proportion of Level 2 units randomized to treatment: J_T / J
ψ_c/ψ_e	0.50	Raito of the total variance of cost data to the total variance of effectiveness data
Parameters for Effectiveness Data		
ρ^e	0.23	Proportion of variance in effectiveness measures that is between clusters
R^2_{1e}	0.50	Proportion of variance of effectiveness data explained by level-1 covariates
R^2_{2e}	0.50	Proportion of variance of effectiveness data explained by level-2 covariates
Parameters for Cost Data		
ρ^c	0.23	Proportion of variance in cost measures that is between clusters
R^2_{1c}	0.00	Proportion of variance of cost data explained by level-1 covariates
R^2_{2c}	0.50	Proportion of variance of cost data explained by level-2 covariates
Parameters for Covariation between Effectiveness Data and Cost Data		
r_1	0.10	Standardized covariance between the effectiveness data and cost data at level-1
r_2	0.10	Standardized covariance between the effectiveness data and cost data at level-2
R^2_{1ec}	0.20	Proportion of the covariance explained by level-1 covariates
R^2_{2ec}	0.50	Proportion of the covariance explained by level-2 covariates
g*	1	Number of Level 2 covariates
n (Average Cluster Size)	50	Mean number of Level 1 units per Level 2 cluster (geometric mean recommended)
J (Sample Size [# of Clusters])	60	Number of Level 2 units
Noncentrality Parameter	**3.02**	Automatically computed from the above assumptions
Power (1-β)	**0.844**	Statistical power (1-probability of a Type II error)

Note: (1)The parameters in yellow cells need to be specified. The power will be calculated automatically. (2) We always assume the effectiveness data are standardized with mean zero and standard deviation one.

Table C6. Demonstration of Geometric Mean Calculation

Cluster #	*Nj*
1	10
2	8
3	14
4	6
5	20
6	9
7	11
8	16
9	13
10	7
11	22
12	15
13	14
14	19
15	17
16	10
17	16
18	4
19	18
20	8
Geometric Mean	11.8