**Title:**

# Advantages of Monte Carlo Confidence Intervals for Incremental Cost-Effectiveness Ratios: A Comparison of Five Methods

**Authors and Affiliations:**

Nianbo Dong*
University of North Carolina at Chapel Hill
116 Peabody Hall, CB 3500
Chapel Hill, NC 27599
Phone: (919)843-9553
dong.nianbo@gmail.com

Rebecca A. Maynard
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
Phone: 609-577-7344
E-mail: rmaynard@upenn.edu

Benjamin Kelcey
University of Cincinnati
3311B RECCENTER
Cincinnati, Ohio 45221
Tel: 513-556-3608
ben.kelcey@gmail.com

Jessaca Spybrook
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
Phone: (269) 387-3889
jessaca.spybrook@wmich.edu

Wei Li
University of Florida
2711R Norman Hall
Gainesville, FL 32611
Phone: 352-273-4332
Email: wei.li@coe.ufl.edu

A. Brooks Bowden
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104

Mobile: 917-940-4649
3700 Walnut Street
Philadelphia, PA 19104
bbowden@upenn.edu

Dung Pham
University of Georgia
College of Agricultural & Environmental Sciences
111 Hoke Smith Building
Athens, Georgia, 30602
Thuy.Pham2@uga.edu

*Corresponding Author

This project is supported by the National Science Foundation [2000705].
The opinions expressed herein are those of the authors and not the funding agency.

Suggested citation:

Dong, N., Maynard, R. A., Kelcey, B., Spybrook, J., Li, W., Bowden, A. B., & *Pham, D. (2024). Advantages of Monte Carlo Confidence Intervals for Incremental Cost-Effectiveness Ratios: A Comparison of Five Methods. *Journal of Research on Educational Effectiveness*. Advance online publication. DOI: 10.1080/19345747.2024.2393412

# Advantages of Monte Carlo Confidence Intervals for Incremental Cost-Effectiveness Ratios: A Comparison of Five Methods

## Abstract

Cost-effectiveness analysis studies in education often prioritize descriptive statistics of cost-effectiveness measures, such as the point estimate of the incremental cost-effectiveness ratio (ICER), while neglecting inferential statistics like confidence intervals (CIs). Without CIs, it becomes impossible to make meaningful comparisons of alternative educational strategies, as there is no basis for assessing the uncertainty of point estimates or the plausible range of ICERs. This study is designed to evaluate the relative performance of five methods of constructing CIs for ICERs in randomized controlled trials with cost-effectiveness analyses. We found that the Monte Carlo interval method based on summary statistics consistently performed well regarding coverage, width, and symmetry. It yielded estimates comparable to the percentile bootstrap method across multiple scenarios. In contrast, Fieller's method did not work well with small sample sizes and treatment effects. Further, Taylor's method and the Box method performed least well. We discussed two-sided and one-sided hypothesis testing based on ICER CIs, developed tools for calculating these ICER CIs, and demonstrated the calculation using an empirical example. We concluded with suggestions for applications and extensions of this work.

*Keywords:*

Confidence interval, Cost-effectiveness analysis (CEA), Fieller's method, Incremental cost-effectiveness ratio (ICER), Monte Carlo interval

# Advantages of Monte Carlo Confidence Intervals for Incremental Cost-Effectiveness Ratios: A Comparison of Five Methods

Cost-effectiveness analysis (CEA) is a type of economic evaluation that compares the costs and effects of alternative programs to identify which program is more efficient at improving an outcome of interest (Levin, et al., 2018; O'Brien et al., 1994; Wakker & Klaassen, 1995; Willan & Briggs, 2006). The comparison is made between each program's incremental cost-effectiveness ratio (ICER), where the costs to deliver the program that produced the effects are divided by the effectiveness estimate. The resulting metric from this ratio is the cost to produce a one unit increase in the outcome. Each ratio is then compared and the program with the lowest cost-effectiveness ratio is preferred. A limitation of CEA studies in education is the lack of precision in reporting ICERs to determine which approach is most efficient. It is difficult to say with confidence if one approach is more efficient than another without additional information on the variation in costs and the resulting variation in the ICER.

Examples of evaluations that include both effects and costs have been increasingly common in education (e.g., Barrett et al., 2020; Bowden & Belfield, 2015; Jacob et al., 2016; Unlu et al., 2015). In the 2024 Request for Applications, the Institute of Education Sciences (IES) recommends understanding the total and incremental costs of the program for strong applications and requires CEA in the research plan for the impact grants (IES, 2024). We anticipate in the future there will be even more educational evaluations that include a cost-effectiveness component, where the costs to produce the effects are reported. One goal of this effort is to strengthen the evidence base on program effects and to support future comparisons among program alternatives where effects are considered relative to their costs. Thus, it is important to examine how the evidence on the costs of interventions is reported and how this can be strengthened to support comparative analyses.

To date, studies that report the costs to produce effects have rarely provided information on how costs vary to support the use of statistical inference to determine if cost-effectiveness ratios are statistically different from one another. In a systematic review of randomized and quasi-experimental studies, we identified 13 publications that reported empirical estimates of the effectiveness and costs of educational interventions, and of those, ten only reported the point estimates of ICERs (Barret et al., 2020; Barrett & VanDerHeyden, 2020; Borman & Hewes, 2002; Bowden et al., 2017; Clark et al., 2020; Finster et al., 2023; Guryan et al., 2020; Hollands et al., 2016; Kim et al., 2011; Scammacca et al., 2020). Only three reported the confidence interval of the ratio (Bowden & Belfield, 2015; Cil et al., 2023; Hunter et al., 2018).

Furthermore, although all 13 publications reported sufficient details about the effectiveness outcomes (e.g., point estimates, standard errors, and *p*-values), none reported standard errors of the incremental costs or the correlations of the incremental costs and effectiveness outcomes. However, relevant cost data may have been collected and more results about incremental costs can be reported, especially for multisite studies. For instance, the cost data were collected at the site level in a multisite randomized trial to evaluate the effectiveness and cost of a curriculum for kindergarten, Zoology One (Gray et al, 2021). The published paper could but did not report the standard error of the incremental cost estimate or the correlation between incremental cost and effectiveness measures. (See the supplemental material for details of the search strategy.)

The failure to report variance and confidence intervals for cost estimates limits the usefulness of study findings since it is difficult to make meaningful comparisons of findings across studies or population groups within studies without knowledge of the precision of the point estimates and plausible range of ICERs. For instance, suppose the reported ICER for Program A was $100 per standard deviation increase in math achievement, while the reported ICER for Program B was $200 per standard deviation increase in math achievement. Comparing the ratios from the two studies alone does not allow for a meaningful judgment about whether Program A is in fact more cost-effective than Program B, since we have no information about the precision of the ratios. These constraints are due, in part, to the difficulty of conducting statistical inference analysis for ratio statistics (e.g., Bowden & Belfield, 2015; Hollands et al., 2013; Levin et al., 2012; Li et al., 2023; O'Brien et al., 1994).

Other fields, including medicine, have overcome these challenges by defining precision levels and calculating and reporting confidence intervals for ICERs. Several approaches to calculating the confidence intervals for ICERs have been developed. For example, Polsky et al. (1997) compared the performance of four methods (the box method, the Taylor series method, the nonparametric bootstrap method, and the Fieller's theorem method) using a Monte Carlo experiment. They found that the bootstrap method and Fieller's theorem method were more accurate than the others in terms of miscoverage rates and symmetrical miscoverage. Polsky et al. (1997) used a binary outcome measure (mortality) to measure the effectiveness (the percentage difference in mortality between the treatment and control groups as the incremental effectiveness measure), did not vary the sample size in their simulations ($n$ = 500), and only tested the percentile bootstrap method. However, in educational studies, continuous outcome measures, such as student academic achievement, are common, and the standardized mean differences as effect sizes are frequently used for the incremental effectiveness measures. In addition, the sample size also affects the coverage rate of the bootstrap methods (Preacher &

Selig, 2012), and the other bootstrap methods (e.g., bias-corrected bootstrap, and bias-corrected and accelerated bootstrap) may have advantages over the percentile bootstrap method when the distribution of the ratio is skewed (Polsky et al., 1997).

Furthermore, the Monte Carlo confidence interval method based on summary statistics has been proposed for testing the product of two parameter estimates, e.g., mediation effects (Preacher & Selig, 2012; Kelcey et al. 2017, 2020), and demonstrated comparable performance with the bootstrap method in constructing confidence intervals in terms of coverage, symmetry, width, and speed.

Monte Carlo confidence intervals have been used to examine the sensitivity (robustness) of cost-effectiveness ratios by changing some assumptions such as the discount rate and the intervention dosage (Boardman et al., 2018, Levin & Belfield, 2015). Also, the Monte Carlo confidence interval method based on the summary statistics (e.g., point estimates and standard errors of the incremental cost and effectiveness, and their correlation/covariance) has been used to demonstrate the problems with Taylor's approximation in calculating confidence intervals for ICERs in a simple simulation (Mullahy & Manning, 1995).

However, the Monte Carlo confidence interval method based on the summary statistics has not been systematically evaluated against all the other methods in constructing confidence intervals for ICERs. In addition, the conventional Monte Carlo confidence interval is constructed using the percentile method (e.g., the 2.5th percentile and the 97.5th percentile of the empirical distribution of the parameter of interest serve as the lower and upper limit of the 95% interval). Just like the bootstrap method, the percentile method may work less well than the bias-corrected Monte Carlo interval for parameters with skewed distributions.

Given the importance of confidence intervals (CIs) for sound application of findings from impact evaluations, the *Standards for the Economic Evaluation of Educational and Social Programs* has suggested that "CEA would ideally include confidence intervals for the incremental cost estimate(s) and the resulting cost-effectiveness ratios" "when an adequate number of cost estimates is available" (Cost Analysis Standards Project, 2021, p.43). Given the lack of precedent and complexity in estimating confidence intervals for ICERs in education studies, it is important to provide statistical tools and guidance for calculating them.

The purpose of this study is twofold. One is to conduct a Monte Carlo experiment to evaluate five methods of constructing CIs for ICERs in randomized controlled trials with cost-effectiveness analyses: (1) the box method, (2) the Taylor series method, (3) the bootstrap method (percentile, bias-corrected, bias-corrected and accelerated), (4) Fieller's theorem method, and (5) the Monte Carlo confidence interval based on the summary statistics (percentile and bias-corrected). We also examine how sensitive the Monte Carlo intervals are to the misspecification of the correlation between costs and effects. The second purpose is to provide statistical tools, which include a SAS macro and a Microsoft Excel-based software, that facilitate accurate computation of CIs of ICERs. We hope this methodological paper will contribute to enhancing the reporting quality of applied research of CEA by providing more detailed information about the incremental costs and the ICER CIs.

In what follows, we first introduce five methods for computing confidence intervals for the ICER. We then describe the procedure for the Monte Carlo experiment and criteria for assessing the adequacy of a confidence interval and present the simulation results. We then discuss the application of ICER CIs for hypothesis testing and demonstrate the calculation of confidence intervals based on two well-performing methods (Fieller's theorem method and the percentile and bias-corrected Monte Carlo confidence intervals based on the summary statistics) using the tools we developed. We use an example from a multisite randomized trial to evaluate the effectiveness and cost of a curriculum for kindergarten, Zoology One (Gray et al, 2021). Finally, we conclude with suggestions and directions for future research.

## Five Methods for Computing Confidence Intervals for ICER

The incremental cost-effectiveness ratio (ICER) is defined as the incremental cost ($\Delta C$) divided by the incremental effectiveness ($\Delta E$): ICER = $\Delta C/\Delta E$ (Bowden et al., 2017; Levin & Belfield, 2015; O'Brien et al., 1994; Wakker & Klaassen, 1995). In educational evaluations, the effectiveness measure for a continuous outcome variable (e.g., math achievement) is usually expressed as an effect size in standard deviation units. Thus, the ICER can frequently be interpreted as the cost per standard deviation increase in the outcome (Bowden et al., 2018; Cost Analysis Standards Project, 2021; Hollands et al., 2016; IES, 2020).

In a randomized trial with a cost-effectiveness analysis with balanced design ($n_1$=$n_0$=$n/2$), we can estimate $\widehat{\Delta C}$, $\widehat{\Delta E}$, and their standard errors ($SE_{\widehat{\Delta C}}$ and $SE_{\widehat{\Delta E}}$) using multivariate regression models and calculated the $\widehat{ICER} = \frac{\widehat{\Delta C}}{\widehat{\Delta E}}$.

$$E_i = \gamma_0^e + \gamma_1^e T_i + \varepsilon_i^e, \quad (1)$$

$$C_i = \gamma_0^c + \gamma_1^c T_i + \varepsilon_i^c, \quad (2)$$

and $$\begin{bmatrix} \varepsilon_i^e \\ \varepsilon_i^c \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & \sigma_{ec} \\ \sigma_{ec} & \sigma_c^2 \end{bmatrix}\right), \quad (3)$$

where $E_i$ and $C_i$ are the effectiveness measure (e.g., test scores) and the cost for participant $i$; $T_i$ is a binary treatment indicator variable ($T_i = 1$ for treatment and 0 for control); $\varepsilon_i^e$ and $\varepsilon_i^c$ are the error terms for effectiveness and cost data, respectively. We assume the error terms follow bivariate normal distributions as shown in equation (3). $\sigma_e^2$ and $\sigma_c^2$ are the variances for the effectiveness and cost error terms and $\sigma_{ec}$ is their covariance. The estimated $\gamma_1^e$ (denoted as $\hat{\gamma}_1^e$ ) and $\gamma_1^c$ (denoted as $\hat{\gamma}_1^c$) represent $\widehat{\Delta E}$ and $\widehat{\Delta C}$, respectively.

Essentially, the parameters can be estimated using the following formulas (Briggs et al., 2002):

$$\widehat{\Delta E} = \hat{\gamma}_1^e = \bar{E}_1 - \bar{E}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1}(E_i\,|T_i = 1) - \frac{1}{n_0}\sum_{i=n_1+1}^{n}(E_i\,|T_i = 0)$$

$$\widehat{\Delta C} = \hat{\gamma}_1^c = \bar{C}_1 - \bar{C}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1}(C_i\,|T_i = 1) - \frac{1}{n_0}\sum_{i=n_1+1}^{n}(C_i\,|T_i = 0)$$

$$Var\big(\widehat{\Delta E}\big) = (SE_{\widehat{\Delta E}})^2 = Var(\bar{E}_1) + Var(\bar{E}_0) = \frac{Var(E_i|T_i = 1)}{n_1} + \frac{Var(E_i|T_i = 0)}{n_0}$$

$$= Var(\hat{\gamma}_1^e) = \frac{\hat{\sigma}_e^2}{n\sigma_T^2} = \frac{4\hat{\sigma}_e^2}{n}$$

$$Var\big(\widehat{\Delta C}\big) = (SE_{\widehat{\Delta C}})^2 = Var(\bar{C}_1) + Var(\bar{C}_0) = \frac{Var(C_i|T_i = 1)}{n_1} + \frac{Var(C_i|T_i = 0)}{n_0}$$

$$= Var(\hat{\gamma}_1^c) = \frac{\hat{\sigma}_c^2}{n\sigma_T^2} = \frac{4\hat{\sigma}_c^2}{n}$$

$$Cov\big(\widehat{\Delta C}, \widehat{\Delta E}\big) = Cov(\bar{C}_1, \bar{E}_1) + Cov(\bar{C}_0, \bar{E}_0)$$

$$= r_1\sqrt{Var(\bar{C}_1)Var(\bar{E}_1)} + r_0\sqrt{Var(\bar{C}_0)Var(\bar{E}_0)} = 2(r_1 + r_0)\frac{\hat{\sigma}_c\ \hat{\sigma}_e}{n}$$

$$= \frac{\hat{\sigma}_{ec}}{n\sigma_T^2} = \frac{r_{\varepsilon^e,\varepsilon^c}\hat{\sigma}_c\hat{\sigma}_e}{n\sigma_T^2} = 4r_{\varepsilon^e,\varepsilon^c}\frac{\hat{\sigma}_c\ \hat{\sigma}_e}{n}$$

$$= r_{\widehat{\Delta C},\widehat{\Delta E}}\sqrt{Var\big(\widehat{\Delta C}\big)Var\big(\widehat{\Delta E}\big)} = 4r_{\widehat{\Delta C},\widehat{\Delta E}}\frac{\hat{\sigma}_c\ \hat{\sigma}_e}{n}$$

Hence, $r_{\widehat{\Delta C},\widehat{\Delta E}} = \frac{(r_1+r_0)}{2} = r_{\varepsilon^e,\varepsilon^c}$, where $r_{\varepsilon^e,\varepsilon^c}$ is the correlation coefficient of the error terms for effectiveness and cost data, and $r_1$ and $r_0$ are the correlation coefficients between costs and effects in the treamtent and control groups, respectively (Briggs et al., 2002).

The $\mathit{\Delta E} - \mathit{\Delta C}$ plane (Figure 1) can be used to facilitate the interpretation of the ICER (e.g., Anderson et al., 1986; Black, 1990; Li et al., 2020; Polsky et al., 1997), where the horizontal axis represents the incremental effectiveness, and the vertical axis represents the incremental cost. The results ($\mathit{\Delta E}$ and $\mathit{\Delta C}$) of a cost-effectiveness analysis can be denoted as a point in Figure 1 (e.g., the blue dot), where the slope of the ray connecting the origin and the point indicates the ICER [ICER = Tan (θ)]. All points on the ray have the same ICER, and the steeper the slope of this ray, the greater the ICER. The potential results of a cost-effectiveness analysis can fall into one of the four quadrants: (1) in Quadrant I, the treatment is more effective but more costly, and it is cost-effective if and only if the estimated ICER < $k$, where $k$ is the cost-effectiveness ratio of the alternative intervention to be compared; (2) in Quadrant II, the treatment is never deemed cost-effective; (3) in Quadrant III, the treatment is cost-effective if and only if the estimated ICER > $k$; and (4) in Quadrant IV, the treatment is always deemed cost-effective. In summary, a treatment is deemed cost-effective if the estimated ICER lies below the red dotted line that represents the ICER = $k$, i.e., the shaded area.

[Figure 1 about here]

Because the $\mathit{\Delta E}$, $\mathit{\Delta C}$, and ICER reported in a cost-effectiveness analysis are the estimates of the population values from the sample, the precision of these estimates depends on the sample sizes and variances of costs and effectiveness measures. The 100(1 – $\alpha$)% confidence interval is often used to measure the precision for estimates with sample variation, where $\alpha$ is the Type I error rate. For instance, a 95% confidence interval, where $\alpha$ = 0.05, defines a range which would include (or "cover") the true population value 95% of the time if the study was repeated an infinite number of times. It also means that the true value would fall outside the interval (i.e., “miscoverage”) 5% of the time. The coverage rate is the proportion of confidence intervals constructed by one method that cover the true population value. Confidence intervals that either over- or under-cover are poorly specified. If the coverage rate of the 95% confidence interval constructed by one method is greater than 0.95 (i.e., over-covering), it expresses too little confidence in the estimate and a narrower confidence interval that provides 95% coverage should be identified; if the coverage rate of the 95% confidence interval is smaller than 0.95 (i.e., under-covering), it expresses too much confidence in the estimate and a wider confidence interval that provides 95% coverage should be identified (Polsky et al., 1997). Ideally, miscoverage of a two-sided 95% confidence interval should be symmetric; that is, the true population value should be smaller than the lower limit of the 95% confidence interval 2.5% of the time and greater than the upper limit of the 95% confidence interval 2.5% of the time.

The formulas for computing the two-sided 100(1 – $\alpha$)% confidence intervals for the $\mathit{\Delta E}$ and $\mathit{\Delta C}$ are readily available, e.g., $CI_{1-\alpha}(\mathit{\Delta E}) = \widehat{\mathit{\Delta E}} \mp t_{1-\alpha} \times SE_{\widehat{\mathit{\Delta E}}}$, where $\widehat{\mathit{\Delta E}}$ is the incremental effectiveness estimate, $t_{1-\alpha}$ is the two-sided critical value of the corresponding Student’s $t$ distribution, and $SE_{\widehat{\mathit{\Delta E}}}$ is the standard error of the $\widehat{\mathit{\Delta E}}$. Because the estimates of the standard errors of the $\mathit{\Delta E}$ and $\mathit{\Delta C}$ are unbiased and efficient when the sample size is sufficiently large and the distributions are approximately normal, the confidence intervals are reliable. However, because the distribution of the ratio may not be well behaved[1] and there is no known unbiased and efficient estimator of the ratio’s standard error, there is no direct method for computing the confidence interval for ICER (O’Brien et al., 1994; Polsky et al., 1997; van Hout et al., 1994; Wakker & Klaassen, 1995).

[1] For example, when the ΔE and ΔC are independent normally distributed variables, the ratio follows a Cauchy distribution. The Cauchy distributed variables can take very extreme values and means of Cauchy distributions may not exist (Wakker & Klaassen, 1995).

Several methods have been proposed to calculate confidence intervals for ICER. Four methods that Polsky et al. (1997) evaluated in a Monte Carlo experiment include: the box methods, the Taylor method, the percentile bootstrap method, and Fieller's theorem method. We evaluate all these four methods. In addition to the percentile bootstrap, we introduce the bias-corrected bootstrap and bias-corrected and accelerated bootstraps below. Furthermore, we examine a new method, i.e., the Monte Carlo interval based on summary statistics, in constructing confidence intervals.

*Box method*

Wakker and Klaassen (1995) described the box method for constructing two-sided confidence intervals for ICER based on Bonferroni's inequality. The box method does not make assumptions about the normality or symmetry of the distributions of $\Delta E$ and $\Delta C$ and it ignores the cost-effectiveness correlation. The confidence interval for the ICER is calculated using confidence limits computed separately for $\Delta E$ and $\Delta C$: $CI_{1-\alpha}(ICER) = \left(\frac{L_C}{U_E}, \frac{U_C}{L_E}\right)$, where $L_E$ and $U_E$ are lower and upper confidence limits for $\Delta E$, and $L_C$ and $U_C$ are lower and upper confidence limits for $\Delta C$, if all four limits are positive; $CI_{1-\alpha}(ICER) = \left(\frac{U_C}{L_E}, \frac{L_C}{U_E}\right)$ if all four limits are negative. The 95% confidence interval for the ICER constructed by the box method using the 95% confidence intervals for the $\Delta E$ and $\Delta C$ is conservative and has been found inappropriately wide (O'Brien et al., 1994; Mullahy & Manning, 1995). To avoid this problem, Polsky et al (1997) proposed to use narrower confidence intervals for the $\Delta E$ and $\Delta C$ to construct the 95% confidence interval for the ICER, that is, using the 68.4% confidence intervals for the $\Delta E$ and $\Delta C$ ($CI_{0.68}(\Delta E) = \widehat{\Delta E} \mp 1.0 \times SE_{\widehat{\Delta E}}$; $CI_{0.68}(\Delta C) = \widehat{\Delta C} \mp 1.0 \times SE_{\widehat{\Delta C}}$) rather than the 95% confidence intervals for the $\Delta E$ and $\Delta C$ ($CI_{0.95}(\Delta E) = \widehat{\Delta E} \mp 1.96 \times SE_{\widehat{\Delta E}}$; $CI_{0.95}(\Delta C) = \widehat{\Delta C} \mp 1.96 \times SE_{\widehat{\Delta C}}$).

*Taylor series method*

O'Brien et al. (1994) applied the delta method, which involves a first-order Taylor series expansion to estimate the standard error of the ICER. The two-sided 95% confidence interval is defined by: $CI_{0.95}(ICER) = \widehat{ICER} \mp 1.96 \times SE_{\widehat{ICER}}$, where $\widehat{ICER} = \frac{\widehat{\Delta C}}{\widehat{\Delta E}}$ and $SE_{\widehat{ICER}} = \sqrt{\left(\frac{\widehat{\Delta C}}{\widehat{\Delta E}}\right)^2 \left(\frac{\left(SE_{\widehat{\Delta C}}\right)^2}{(\widehat{\Delta C})^2} + \frac{\left(SE_{\widehat{\Delta E}}\right)^2}{(\widehat{\Delta E})^2} - \frac{2Cov(\widehat{\Delta C},\widehat{\Delta E})}{\widehat{\Delta C}\widehat{\Delta E}}\right)}$. This method assumes a normal distribution for the ICER and incorporates the cost-effectiveness correlation into its standard error calculation.

*Bootstrap method*

Bootstrapping method (Efron, 1979; Efron & Tibshirani, 1993) is a nonparametric method that involves resampling from the study sample, computing the ICER in each of multiple samples, and constructing the confidence interval from the empirical sampling distribution of the ICERs (Chaudhary & Stearns, 1996; O'Brien et al., 1994). The first step of the procedure is to independently draw an arbitrarily large number (*B*) of resamples of size *N* with replacement from the original sample. Then the $\Delta E$, $\Delta C$, and ICER are calculated for each of these *B* resamples, resulting in an empirical sampling distribution of ICER. Then the confidence intervals are constructed, which differ for particular bootstrap approaches as defined next.

The percentile bootstrap (Efron, 1981) uses the lower and upper $50\alpha$% of the distribution of the estimated ICER to define the two-sided 100(1-$\alpha$)% confidence interval of ICER. The percentile bootstrap is simple to compute but may not work well if the bootstrap distribution is asymmetric. Efron (1981, 1982, 1987) and Efron and Tibshirani (1993) proposed the bias-corrected bootstrap to reduce bias by incorporating an adjustment. Letting $z_0$ be the *z*-score corresponding to the proportion of the *B* bootstrap resamples with the estimated ICER from each resample less than the estimated ICER from the original sample, two *z*-scores are defined as:

$z'_{lower} = 2z_0 + z_{\alpha/2}$ and $z'_{upper} = 2z_0 + z_{1-\alpha/2}$, where $2z_0$ is a correction for median bias. The proportions under the standard normal distribution that correspond to $z'_{lower}$ and $z'_{upper}$ are multiplied by 100 to serve as the adjusted percentiles for selecting the lower and upper confidence limits from the bootstrap distribution of the estimated ICER. The bias-corrected intervals do not always have good coverage (Schenker, 1985). Efron (1987) proposed the bias-corrected and accelerated confidence intervals, which include further adjustment of skewness in the bootstrap distribution by an acceleration constant $\dot{a}$: $z'_{lower} = z_0 + \frac{z_0+z_{\alpha/2}}{1-\dot{a}(z_0+z_{\alpha/2})}$ and $z'_{upper} = z_0 + \frac{z_0+z_{1-\alpha/2}}{1-\dot{a}(z_0+z_{1-\alpha/2})}$, where $\dot{a}$ is a correction for skewness and is approximately 1/6 of the skewness of the bootstrap distribution of ICER. Note that when $z_0 = 0$, the bias-corrected interval is same as the percentile interval; when $\dot{a} = 0$, the bias-corrected and accelerated interval is same as the bias-corrected interval; and when $z_0 = \dot{a} = 0$, the bias-corrected and accelerated interval is same as the percentile interval. The primary benefits of bootstrapping are that it involves no distributional assumptions on the $\Delta E$, $\Delta C$, and ICER, and it considers the cost-effectiveness correlation.

*Fieller's theorem method*

Fieller (1954) proved a theorem for computing confidence intervals of a ratio by transferring the ratio into a linear function of two variables. The numerator and denominator of the ratio are assumed to follow a bivariate normal distribution. This method has been applied to construct confidence intervals for the ICER (Chaudhary & Stearns, 1996; O'Brien et al., 1994; Willan & O'Brien, 1996).

In this case, we define the incremental net monetary benefit (INMB): $INMB = \kappa\widehat{\Delta E} - \widehat{\Delta C}$, where $\kappa$ is the "willingness to pay" (Willan & Briggs, 2006), as the threshold ICER that renders the intervention cost-effective (Stinnett & Mullahy, 1998). Then, observe $\frac{\kappa\widehat{\Delta E}-\widehat{\Delta C}}{\sqrt{\left(\kappa^2\left(SE_{\widehat{\Delta E}}\right)^2+\left(SE_{\widehat{\Delta C}}\right)^2-2\kappa Cov\left(\widehat{\Delta C},\widehat{\Delta E}\right)\right)}} \sim N(0,1)$. The two-sided confidence interval at a significance level, $\alpha$, is obtained by solving the inequality: $\left|\frac{\kappa\widehat{\Delta E}-\widehat{\Delta C}}{\sqrt{\left(\kappa^2\left(SE_{\widehat{\Delta E}}\right)^2+\left(SE_{\widehat{\Delta C}}\right)^2-2\kappa Cov\left(\widehat{\Delta C},\widehat{\Delta E}\right)\right)}}\right| \leq z_{\alpha/2}$. That is, $a\kappa^2 + b\kappa + c \leq 0$, which is called the Fieller quadratic, where $a = \left(\widehat{\Delta E}\right)^2 - z_{\alpha/2}^2(SE_{\widehat{\Delta E}})^2$, $b = -2\left[\widehat{\Delta C} \cdot \widehat{\Delta E} - z_{\alpha/2}^2 Cov\left(\widehat{\Delta C}, \widehat{\Delta E}\right)\right]$, and $c = \left(\widehat{\Delta C}\right)^2 - z_{\alpha/2}^2(SE_{\widehat{\Delta C}})^2$. When $b^2 \leq 4ac$, there is no real solution for the inequality; when $b^2 > 4ac$ and $a < 0$ (i.e., $\left(\frac{\widehat{\Delta E}}{SE_{\widehat{\Delta E}}}\right)^2 < z_{\alpha/2}^2$), the solution is $(-\infty, \min\{l_1, l_2\}) \cup (\max\{l_1, l_2\}, +\infty)$, where $l_{1,2} = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$; when $b^2 > 4ac$ and $a > 0$ (i.e., $\left(\frac{\widehat{\Delta E}}{SE_{\widehat{\Delta E}}}\right)^2 > z_{\alpha/2}^2$), the solution is $(\min\{l_1, l_2\}, \max\{l_1, l_2\})$. In particular, when $b^2 > 4ac$ and $a$ is close to 0 (i.e., $\left(\frac{\widehat{\Delta E}}{SE_{\widehat{\Delta E}}}\right)^2$ is close to $z_{\alpha/2}^2$), the solution is close to $(-\infty, +\infty)$.

In summary, only when $b^2 > 4ac$ and $a > 0$ (i.e., $\left(\frac{\widehat{\Delta E}}{SE_{\widehat{\Delta E}}}\right)^2 > z_{\alpha/2}^2$, indicating a significant treatment effect estimate on the effectiveness measure), does Fieller's theorem provide a meaningful confidence interval. Fieller's theorem method does not restrict the distribution of the ICER to be normal or symmetric, and it includes the cost-effectiveness correlation in the computation.

*The Monte Carlo confidence interval based on the summary statistics*

The Monte Carlo confidence intervals have been advocated for the sensitivity (robustness) analysis of the cost-effectiveness ratios by changing some assumptions such as the discount rate and the intervention dosage (Boardman et al., 2018). Mullahy and Manning (1995) also use the Monte Carlo confidence interval method based on the summary statistics (point estimates and standard errors of the incremental cost and effectiveness measures, and their correlations/covariances) to demonstrate the problems of Taylor's approximation in calculating confidence intervals for ICER in a simple simulation. However, the Monte Carlo confidence interval method based on the summary statistics has not been systematically evaluated against all the other methods in constructing confidence intervals for ICER. In contrast, the Monte Carlo confidence interval based on summary statistics has been evaluated for constructing confidence intervals for the product of two parameter estimates (i.e., mediation effects) and, in that application, demonstrated comparable performance with the bootstrap method regarding the coverage, width, symmetry, and speed (Preacher & Selig, 2012; Bai et al., 2023; Cox & Kelcey, 2023; Kelcey et al. 2017, 2020). In addition, the conventional two-sided Monte Carlo confidence interval is constructed using the percentile method, e.g., the 2.5th percentile and the 97.5th percentile of the empirical distribution of the parameter of interest serve as the lower and upper limit of the 95% interval. Just like the bootstrap method, the percentile method may work less well when the distribution of the parameter is skewed, whereas the bias-corrected Monte Carlo interval may work better.

The Monte Carlo interval method makes the assumption that the parameters $\mathit{\Delta E}$ and $\mathit{\Delta C}$ have a joint normal sampling distribution, with parameters supplied by the estimates on the $\mathit{\Delta E}$ and $\mathit{\Delta C}$: $\begin{bmatrix}\Delta C^{*}\\ \Delta E^{*}\end{bmatrix} \sim MVN\left(\begin{bmatrix}\widehat{\mathit{\Delta C}}\\ \widehat{\mathit{\Delta E}}\end{bmatrix}, \begin{bmatrix}(SE_{\widehat{\mathit{\Delta C}}})^2 & Cov(\widehat{\mathit{\Delta C}},\widehat{\mathit{\Delta E}})\\ Cov(\widehat{\mathit{\Delta C}},\widehat{\mathit{\Delta E}}) & (SE_{\widehat{\mathit{\Delta E}}})^2\end{bmatrix}\right)$. A sampling distribution of $\frac{\widehat{\mathit{\Delta C}}}{\widehat{\mathit{\Delta E}}}$ can be formed by repeatedly generating $\Delta C^{*}$ and $\Delta C^{*}$ and computing their ratio many times (e.g., $M$ = 100,000). Although parametric assumptions are invoked for $\widehat{\mathit{\Delta C}}$ and $\widehat{\mathit{\Delta E}}$, no parametric assumptions are made about the distribution of $\frac{\widehat{\mathit{\Delta C}}}{\widehat{\mathit{\Delta E}}}$. Percentiles for this sampling distribution are identified to serve as the limits for a 100(1-$\alpha$)% confidence interval of ICER.

Similar to the bias-corrected bootstrap method, the bias-corrected Monte Carlo interval can be constructed as follows. Letting $z_0$ be the *z*-score corresponding to the proportion of the *M* parameter sets with the estimated ICER ($\frac{\Delta C^*}{\Delta E^*}$) from each parameter set less than the estimated ICER ($\frac{\widehat{\Delta C}}{\widehat{\Delta E}}$) from the original sample, two *z*-scores are defined as: $z'_{lower} = 2z_0 + z_{\alpha/2}$ and $z'_{upper} = 2z_0 + z_{1-\alpha/2}$, where $2z_0$ is a correction for median bias. The proportions under the standard normal distribution that correspond to $z'_{lower}$ and $z'_{upper}$ are multiplied by 100 serves as the adjusted percentiles for selecting the lower and upper confidence limits from the sampling distribution of $\frac{\widehat{\Delta C}}{\widehat{\Delta E}}$.

## The Monte Carlo Experiment

We conducted a Monte Carlo experiment to evaluate the performance of five methods for constructing two-sided confidence intervals for ICERs. Specifically, these methods include: the box method (1.96 SE and 1.0 SE), Taylor method, the bootstrap method (percentile, bias-corrected, and bias-corrected and accelerated), Fieller's theorem method, and the Monte Carlo intervals based on the summary statistics (percentile and bias-corrected) (Table 1). In addition, we examined the confidence intervals using the Monte Carlo interval method with misspecified cost-effectiveness correlations (assuming the correlation to be 0 while the true correlation is non-zero).

### *Procedure*

The procedures are below:

(1) We generated cost and effectiveness data for randomized trials. We varied the sample sizes, the distributions of the cost, the correlations between the cost and the effectiveness measures, and incremental effectiveness ($\Delta E$). The sample sizes (*n*) were 20, 40, 60, 100, 150,

200, 300, 400, 500, 600, and 800. The sample was randomly assigned to either the treatment group or control group in each trial. By definition, the ICER is a linear function of $\Delta C$, but it is not a monotone function of $\Delta E$. When $\Delta E$ is close to 0, the ICER approaches infinity, which implies that when $\widehat{\Delta E}$ is not statistically different from 0, there is a chance the ICER estimate will be infinity and the distribution of the ICER estimate can be bimodal— conditions that affect the confidence intervals. By following Polsky et al.'s (1997) simulation, we used a significant $\widehat{\Delta C}$, but we allowed $\widehat{\Delta E}$ to be both significant and non-significant in our simulations to investigate how the significance of $\widehat{\Delta E}$ affected the confidence intervals; in addition, we used both normal and lognormal distributions for the costs. Thus, we assumed the cost estimates have a normal distribution with a mean of \$20,000 and standard deviation (SD) of 2,000 for the entire sample, and we added $\Delta C$ = \$5,000 to the cost for the treatment group; We also assumed the cost estimates have a lognormal distribution with a mean of \$20,000 and the SD of 8,000 for the entire group, and we added $\Delta C$ = \$5,000 to the cost for the treatment group. We assumed the effectiveness outcome has a normal distribution with a mean of 1.0 and the SD of 1.0 for the entire sample, and we added $\Delta E$ = 0.25 and 0.50, respectively, to the effectiveness outcome for the treatment group.

Hill et al (2008) reported that the effect sizes from 61 randomized studies (468 effect sizes) are in the 0.07-0.51 range and the mean effect sizes from the meta-analysis of 76 meta-analyses of educational interventions are in the 0.20-0.30 range. Recently, Kraft (2020) reported that the mean effect size is 0.16 (SD = 0.28) and the median effect size is 0.10 (the 30th percentile is 0.02 and the 70th percentile is 0.21) from the meta-analysis of 747 randomized studies (1,942 effect sizes). We chose 0.25 to represent the effect size of an effective educational intervention. In addition, we used an effect size of 0.50 to evaluate how the effect sizes may affect the performance of various methods. Together with the sample sizes, these effect sizes can provide a wide range of significance levels of educational interventions (*t* = 0.56 – 7.07).

Empirical results about the correlations of the cost and effectiveness estimates in educational research are rarely reported. We used data from a multisite study (Gray et al, 2021) to get an empirical estimate of the correlation (0.33) among 19 sites. We followed Polsky et al.'s (1997) simulation and used a wide range of correlations from -0.5 to 0.5 in a step of 0.1 in the simulation. This range of correlations can allow us to explore how the correlations affect the performance of various methods. This resulted in a total 484 scenarios (11 sample sizes × 2 cost distributions × 2 $\Delta E$ × 11 correlations).

(2) For each scenario, we estimated $\widehat{\Delta C}$, $\widehat{\Delta E}$, and their standard errors ($SE_{\widehat{\Delta C}}$ and $SE_{\widehat{\Delta E}}$) using two ordinary least square regression models (Equations 1-3), estimated the cost-effectiveness correlation using $r_{\widehat{\Delta C},\widehat{\Delta E}} = (r_T + r_C)/2$, and calculated the $\widehat{ICER} = \frac{\widehat{\Delta C}}{\widehat{\Delta E}}$. We constructed two-sided 95% confidence intervals for the ICERs using five methods discussed above. Specifically, we used the box method based on both 1.96 standard errors and 1.0 standard errors of $\widehat{\Delta C}$ and $\widehat{\Delta E}$ when four limits are all positive or all negative, and we coded the other situations as "exclusive". We used the Taylor series method based on the standard error of $\widehat{ICER}$ derived from the first-order Taylor series expansion.

For bootstrapping, we used the macros developed by SAS Institute Inc. (2007) with modifications for calculating confidence intervals for the percentile, bias-corrected, and bias-corrected and accelerated methods. We used Fieller's method when $\left(\frac{\widehat{\Delta E}}{SE_{\widehat{\Delta E}}}\right)^2 > z_{\alpha/2}^2$, and coded the other situations as "exclusive". For the Monte Carlo interval based on the summary statistics, we used both the percentile and bias-corrected methods; in addition, we constructed confidence intervals using the misspecified correlation between the cost and effectiveness outcome (always assuming 0).

If the true ICER fell within the two-sided 95% CI, we coded the variable "coverage" as 1 (0 otherwise); if the lower limit of the 95% CI exceeded the true ICER, we coded the variable "left_side_miscoverage" as 1. The interval width was calculated by the difference of the upper and lower confidence limits, i.e., $U_{ICER} - L_{ICER}$.

(3) We conducted 3,000 replications for each scenario. We calculated the mean of the variable "coverage" as the coverage rate (*CR*), the mean of the variable "exclusive" as the exclusion rate, the average width, and the proportion of "left_side_miscoverage" in the total miscoverage as the symmetry measure over 3,000 replications. In addition, we calculated the bias of the coverage rate and the root mean square error (*RMSE*) of the coverage rate across multiple scenarios: $Bias_k = CR_k - 0.95$ and $RMSE = \sqrt{\frac{\sum_{k=1}^{K}(CR_k - 0.95)^2}{K}}$, where $CR_k$ and $Bias_k$ are the coverage rate and bias for the $k^{th}$ scenario. Furthermore, we regressed the absolute value of the bias on the sample size, the $\Delta E$ (0.25 or 0.50), the distribution of the cost (lognormal or normal), and the correlation between the cost and effectiveness to investigate the effects of these factors.

### *Criteria for Evaluating the Methods*

The primary evaluation criterion is the coverage rate, which is preferred to the interval width and symmetry measure for the ratio interval (Jiang et al., 2000) and the product interval (e.g., mediation effect, Preacher & Selig, 2012). The method with the best confidence interval is the one that produces the smallest RMSE of the coverage rate across multiple scenarios. In addition, a method is better if it comes closer to the target coverage rate of 95%. The interval width and symmetry measure are the secondary criteria. The method is better if it produces narrower width and symmetry measure closer to 0.5.

## *Results*

### *Summative Assessment*

Across all 484 simulation scenarios, the ICER estimates over 3,000 replications had an average skewness of 1.58 with a range of -54.77 to 54.77 and an average kurtosis of 1145.88 with a range of 0.58 to 2999.92. The average exclusion rate for Fieller's method was 0.37 with the maximum of 0.91. This means that, on average, 37% of the replications were not able to produce meaningful confidence intervals using Fieller's method. The Box method (1.96 SE) and the Box method (1.0 SE) had average exclusion rates of 0.39 and 0.19 with maximum rates of 0.99 and 0.84, respectively.

The *RMSE* of the coverage rate, the average coverage rate, the average interval width, and the average symmetry across 484 simulation scenarios by the confidence interval constructing method are reported in Table 1. The percentile bootstrap method and the Monte Carlo interval based on the summary statistics produced the smallest *RMSE* (0.018 and 0.019, respectively). Surprisingly, the Monte Carlo interval based on the summary statistics when the correlation between costs and effects was misspecified (always assuming 0) produced a slightly larger *RMSE* (0.024). The box method (1.96 SE) produced the next smallest *RMSE* (0.049) followed by Taylor's method (0.084) and Fieller's method (0.095). The bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, the bias-corrected Monte Carlo interval, and the Box method (1.0 SE) produced similar *RMSE* ranging from 0.114 to 0.119.

[Table 1 about here]

The percentile bootstrap method produced the closest average coverage rate (0.965) to 0.95 along with the Monte Carlo interval (0.966) and the Monte Carlo interval method with a misspecified correlation (0.966). The Box method (1.96 SE) produced the next closest average coverage rate (0.975) followed by Fieller's method (0.914). In addition, the percentile bootstrap method and the Monte Carlo interval produced narrower ranges for the average coverage rate (0.924-0.983 and 0.934-0.983) than did the Monte Carlo interval with a misspecified correlation (0.888-0.997), Fieller's method (0.530-0.980) and the Box method (1.96 SE) (0.715-1.000).

The bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval resulted in under-coverage and produced similar average coverage rates around 0.88 and coverage range0.58-0.98. Both Taylor's method and the Box method (1.0 SE) also resulted in under-coverage, with an average coverage rate of 0.890 with a wide range (0.680-0.960) and an average coverage rate of 0.860 with a wide range (0.523-0.975), respectively.

The percentile bootstrap method, the Monte Carlo interval, and the Monte Carlo interval with a misspecified correlation produced similarly small average coverage width with narrow ranges. All the other methods produced larger widths with larger ranges. The bias-corrected bootstrap method, the bootstrap bias-corrected and accelerated method, and the bias-corrected Monte Carlo interval produced better symmetry with an average of 0.38 and a range of 0.04-0.60 than the other methods.

Overall, ranges on the average coverage rate, width, and symmetry varied considerably across 484 simulation scenarios. Thus, it is important to investigate which simulation factors contribute to the variation. Table 2 presents the results of regressing the absolute value of the bias for the 95% coverage rate on the sample size, the effectiveness effect size ($\Delta E$), the cost-effectiveness correlation, and the distribution of costs (lognormal vs. normal). The proportion of variance on the absolute value of the bias explained by these four simulation factors ranges from 0.02 to 0.65 ($R^2$ = 0.02-0.65) across all the methods.

[Table 2 about here]

All of these simulation factors were mean-centered, hence, the intercepts can be interpreted as the mean absolute value of the bias. The bootstrap percentile method and the Monte Carlo interval produced similarly smaller mean absolute bias (0.016 and0.017) than the other methods. This pattern is similar to the *RMSE* in Table 1. The sample size is statistically significant ($p < 0.01$) across all methods. The $\Delta E$ was statistically significant ($p < 0.01$) across all methods except for the Box CI (1.96 SE) and the Monte Carlo interval with a misspecified correlation. The magnitudes of the associations (coefficients) of the sample size and effectiveness effect size with the absolute value of the bias are smaller for the bootstrap percentile method and the Monte Carlo interval than the other methods.

When the sample size and the effectiveness effect size increased, the absolute value of the bias decreased, except for the Monte Carlo interval with a misspecified correlation. The cost-effectiveness correlation only affected the Box method (1.0 SE) and the Monte Carlo interval with a misspecified correlation. In addition, the quadratica term of the cost-effectiveness correlation also affected the bias for the Monte Carlo interval with a misspecified correlation. The distribution of costs (lognormal or normal) only affected the Box method (1.0 SE), Taylor's method, and the Monte Carlo interval with a misspecifed correlation.

*Summary across 22 scenarios by the sample size and the $\Delta E$*

Given that the sample size and the $\Delta E$ played a more consistently important role in explaining the bias across methods, we presented the summary results of *RMSE*, coverage rate, width, and symmetry by the $\Delta E$ and sample sizes across 22 scenarios (2 cost distributions × 11 correlations) in Figures 2 and 3 and Tables 3 and 4. Based on the assigned treatment effects on the effectiveness measure and sample sizes, we also reported the *t*-statistics in the figure and tables: $t = \frac{\Delta E}{SE_{\Delta E}} = \frac{\Delta E\sqrt{n}}{2}$, where $\Delta E$ = 0.25 or 0.50, *n* = 20 – 800, and the variance of the effectiveness measure is 1 in the simulation. Specifically, Figures 2a and 2b presented the *RMSE* of 95% confidence interval coverage as a function of sample size for $\Delta E$ = 0.25 and 0.50, respectively; Figures 3a and 3b presented the average 95% confidence interval coverage as a function of sample size for $\Delta E$ = 0.25 and 0.50, respectively. In all four figures, the long grey dash line represents the Box method (1.0 SE), the short grey dash line represents the Box method (1.96 SE), the cyan solid line represents Taylor's method, the red solid line represents the Monte Carlo interval based on the summary statistics, the long shadow blue dash line represents the percentile bootstrap method (Bootstrap PCTL), the short green dashed line represents Fieller's method, the short dashed blue line represents the bias-corrected bootstrap method (Bootstrap BC), the orange solid line represents the bias-corrected and accelerated bootstrap method (Bootstrap BCa), and the long dark blue dash line represents the bias-corrected Monte Carlo interval based on the summary statistics (MC corrected).

[Figures 2-3 about here]

[Tables 3-4 about here]

The percentile bootstrap method and the Monte Carlo interval based on the summary statistics had consistently good performance (small *RMSE* and close coverage to 0.95) across $\Delta E$

and sample sizes (and $t$-statistics). When $t \geq 2.5$ (i.e., $\Delta E$ = 0.25 and $n \geq 400$ or $\Delta E$ = 0.5 and $n \geq$ 100), Fieller's method, the bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval had similarly good performance as the bootstrap percentile method and the Monte Carlo interval ($RMSE$ < 0.023 and coverage = 0.942-0.972). However, when the $t$-statistic is small, for example, $t$ < 1.77 (i.e., $\Delta E$ = 0.25 and $n$ < 200 ), the bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval produce larger $RMSE$ (> 0.07) and under coverage rate (< 0.88). When $\Delta E$ = 0.25 and $n$ = 100, i.e., $t$ = 1.25, Fieller's method produced $RMSE$ = 0.07 with an average coverage = 0.881 and the exclusion rate = 0.76, which means that Fieller's method failed to construct the confidence intervals in 76% of the replications. Taylor's method produced consistently under-covering coverage rates and approached good performance only when $t \geq 5.00$ ($\Delta E$ = 0.5 and $n \geq 400$). The Box methods (1.96 SE) always produced over-covering 95% coverage rates (0.956-0.993) except when $t$ < 1.00 and the Box methods (1.0 SE) always produced under-covering 95% coverage rates (0.805-0.902).

The percentile bootstrap method, the Monte Carlo interval, and the Monte Carlo interval with a misspecified correlation produced similarly smallest average coverage width consistently across $\Delta E$ and sample sizes. When $t \geq 5.0$ (e.g., $\Delta E$ = 0.5 and $n \geq 400$), Fieller's method, the bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval had performance on width and symmetry (0.46-0.51) similar to the percentile bootstrap method and the Monte Carlo interval.

The performance of these methods depends on the joint effect of the sample size and the $\Delta E$ (i.e., $t$-statistic) because the ICER is a ratio statistic. By definition, the ICER is not a monotone function of $\Delta E$. When $\Delta E$ is close to 0, the ICER approaches infinity. It implies that when $\widehat{\Delta E}$ is not statistically different from 0 (i.e., the *t*-statistic is smaller than the critical *t*), there is a chance the ICER estimate will be infinity and the distribution of the ICER estimate can be bimodal, which affect the confidence intervals.

*Summary across 44 scenarios by the cost-effectiveness correlation*

To investigate how the misspecification of the cost-effectiveness correlation using the Monte Carlo interval method affects confidence intervals, we present the summary results of *RMSE* and average coverage rates across 44 scenarios (11 sample sizes × 2 cost distributions × 2 $\Delta E$) as a function of the cost-effectiveness correlation (-0.5 to 0.5) in Figures 4 and 5. In addition to the Monte Carlo interval with a misspecified correlation (assuming 0), we also plotted the Monte Carlo interval with correctly specified correlation and the percentile bootstrap method as references, and the Box method (1.96 SE and 1.0 SE) as a comparison. The purple dotted line represents the Monte Carlo interval with a misspecified correlation, and the other methods used the same lines as forementioned. Not surprisingly, the *RMSE* of the 95% coverage rate for the Monte Carlo interval with a misspecified correlation had a curvilinear relationship with the correlation. When the true correlation is within (-0.3, 0.1), the misspecification of correlation has small effects on the *RMSE* and average coverage rate. Even when the correlation is misspecifed, the Monte Carlo interval produced smaller *RMSE* (0.016-0.035) and better coverage rates (0.945-0.984) than the Box method.

[Figures 4 & 5 about here]

In summary, the Monte Carlo interval based on the summary statistics had the same good performance (coverage, width, and symmetry) as the percentile bootstrap method in constructing the two-sided confidence intervals for the ICER across all simulation scenarios. Both methods were robust against the cost-effectiveness correlations and the distributions of costs (lognormal or normal) and produced better coverage, narrower width, and better symmetry when the sample size and the $\Delta E$ increased. Fieller's method, the bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval had good performance when the $\Delta E$ and sample sizes were large (e.g., $\Delta E = 0.5$ and $n \geq 400$, i.e., $t \geq 5.0$). Taylor's method and the Box method (1.0 SE or 1.96 SE) had worse performance than the aforementioned methods. In addition, the misspecification of correlation (assuming 0) using the Monte Carlo interval had small effects on *RMSE* and average coverage rate when the true correlation was within (-0.3, 0.1).

## Application of ICER CIs for Hypothesis Testing and Demonstration of the Software

The descriptive point estimate of the ICER itself is useful, but not enough for hypothesis testing or policy making. The ICER CIs provide a range of values within which the true value of ICER likely lie with confidence. The width of a CI measures the precision of the estimated ICER. Narrower intervals indicate more precise estimates, while wider intervals suggest less precision. This helps in understanding the reliability of the ICER estimates derived from a sample. In addition, the CIs can be used for hypothesis testing and in decision-making processes.

For two-sided hypothesis testing in cost-effectiveness studies, the null hypothesis is typically stated as $H_0$: ICER = $k$, meaning that the true ICER equals a threshold value $k$. The alternative hypothesis for a two-sided test is $H_a$: ICER $\neq k$. For a left-sided test, $H_0$: ICER $\geq k$, meaning that the true ICER is equal or greater than $k$ and tested program is equally or less cost-effective; the alternative hypothesis is $H_a$: ICER $< k$, indicating that the true ICER is less than $k$ and the tested program is more cost-effective. For a right-sided test, $H_0$: ICER $\leq k$, meaning that the true ICER is equal or smaller than $k$ and the tested program is equally or more cost-effective;

$H_a$: ICER > $k$, suggesting that the true ICER is greater than $k$ and the tested program is less cost-effective.

When $\alpha = 0.05$, the 95% confidence interval for two-sided test is the values associated with the 2.5th and 97.5th percentiles. In the case of the left-sided test, the 95% confidence interval's lower and upper limits are -∞ and the value at the 95th percentile, respectively; Conversely, for the right-sided test, these limits are the value at the 5th percentile and +∞, respectively. A hypothesis test is statistically significant when the $100(1 - \alpha)\%$ confidence interval does not include $k$.

We developed a Microsoft Excel-based software and a SAS macro for calculating the two-sided and one-sided confidence intervals of the ICER using the Monte Carlo interval. In addition, we provided two-sided confidence intervals for Fieller's method as a comparison. Unlike the bootstrap method, which requires the original data, these methods need only point estimates and standard errors of the incremental cost and effectiveness measures and their correlation. Once these summary statistics are input, the ICER and their confidence intervals will be automatically calculated for a specified confidence level.

To demonstrate this software, we use an example from the forementioned multisite randomized trial to evaluate the effectiveness and cost of the Zoology One kindergarten curriculum (Gray et al, 2021). This example is a good representation of typical efficacy trials in education that educational outcomes data were measured at the individual level, but cost data were often collected at the site level. This example does not follow the ideal where we have costs per individual but we chose to use this example intentionally to demonstrate the wide and easy application of our methods. The original article (Gray et al, 2021) reported neither the standard error for the incremental cost estimate nor the correlation between incremental cost and effectiveness measures. However, it did estimate the incremental costs per student using the ingredient method for 20 out of 21 schools in the original multisite study. In addition, the treatment effect sizes and their standard errors on the Woodcock Reading Mastery Test passage comprehension were estimated for 20 out of 21 schools.

The study reported both incremental costs and effectiveness outcome estimates for 19 schools. Thus, to recover these estimates for constructing ICER CIs, we applied methods typically used in Meta-Analysis to compute the point estimate and standard error of the incremental effectiveness of the tested intervention. We then calculated the mean and standard error of the incremental cost and the correlation of the incremental effectiveness and cost for these 19 schools. The estimated average effect size across these schools on the Woodcock Reading Mastery Test passage comprehension is 0.15 (SE = 0.04) and the incremental cost per student is $499.36 (SE = 48.91). The correlation between the incremental cost and effectiveness across 19 schools is 0.33.

After these parameters are entered into the software and a desired confidence interval (95% in this case) is specified, the software produces an ICER estimate and its two-sided and one-sided 95% confidence intervals.

Figure 6 presents the screenshot of this calculation based on the Excel-based software. The estimated ICER is $3,293, that is, the study findings suggest that, on average, using Zoology One, it would cost an average of $3,293 per pupil to increase passage comprehension scores by one standard deviation. The two-sided 95% confidence intervals of the ICER are ($2,209, $6,665) based on the Monte Carlo interval and ($2,209, $6,648) based on Fieller’s method. For the one-sided test, the left-sided 95% CI is (-∞, 5,721) and the right-sided 95% CI is (2,338, ∞). Figures S1 and S2 in the supplemental material present similar results using the SAS macro.

The distribution of the ICER based on Monte Carlo simulation is right-skewed. Note that the confidence intervals calculated from the Monte Carlo method change every time after the simulation is rerun. Increasing the number of replications of simulation will reduce the change. In this example, the 95% confidence intervals are very close between the Monte Carlo interval and Fieller's method, providing greater confidence in the results.

The two-sided 95% confidence interval of the ICER based on the Monte Carlo interval suggests that, if the studies are replicated infinitely, the true cost per student for each standard deviation increase on passage comprehension falls between $2,209 and $6,665 95% of times. If we knew that the true ICER for the alternative Program X was within this CI, we would not reject the null hypothesis that Zoology One was equally cost-effective as Program X. On the other hand, If we knew that the true ICER for the alternative Program X was outside of this CI, we could determine that Zoology One was not as cost-effective as Program X.

If the true ICER for the alternative Program X was below the lower limit of the right-sided 95% CI (i.e., $2,338), we would conclude that Zoology One was less cost-effective than Program X based on one-sided testing; If the true ICER for the alternative Program X was above the upper limit of the left-sided 95% CI (i.e., $5,721), we would conclude that Zoology One was more cost-effective than Program X based on one-sided testing.

These ICER CIs and hypothesis testing can help stakeholders understand the range of potential ICERs and the uncertainty associated with those ICER estimates and assist with policy making, for example, choosing the statistically more cost-effective programs.

[Figure 6 about here]

# Conclusion

Through a Monte Carlo experiment, we have found that the Monte Carlo interval based on summary statistics has consistently good performance regarding coverage, width, and symmetry as the percentile bootstrap method in constructing the confidence intervals for the ICER across multiple scenarios (varying sample sizes, $\Delta E$, cost-effectiveness correlations, and distributions of costs). Although Fieller's method has been recommended in the health literature (e.g., Briggs et al., 1999; Chaudhary & Stearns, 1996; Polsky et al., 1997) and the bias-corrected bootstrap method has also been recommended by Chaudhary and Stearns (1996), our study found that these two methods did not work well when the sample size and the effectiveness treatment effect size were small. Taylor's method and the Box method (1.0 SE or 1.96 SE) had worse performance than the Monte Carlo interval and the percentile bootstrap method. The poor performance of Taylor's method is consistent with the literature (e.g., Mullahy & Manning, 1995). Thus, we suggest using the percentile bootstrap method and the Monte Carlo interval based on the summary statistics to construct the confidence intervals of the ICER. When the *t*-statistic of the effectiveness measure is larger than 2.5, Fieller's method, the bias-corrected bootstrap method, the bias-corrected and accelerated bootstrap method, and the bias-corrected Monte Carlo interval can also be used.

One advantage of the Monte Carlo interval is that it is easier to compute compared to other methods (e.g., the bootstrap method), especially when the data structure of the cost and effectiveness measures are complicated. The Monte Carlo interval relies on the summary statistics (point estimates and standard errors of the incremental cost and effectiveness, and their correlation), which can be estimated from the conventional multilevel or ordinary least square analysis of the costs and effectiveness, while the bootstrap method may involve resampling the clusters of participants. For example, in multisite cost-effectiveness studies, the effectiveness data are collected at the individual level and the cost data are collected at the site level, which are very common in educational evaluations. Researchers can first estimate the incremental costs and effectiveness outcome by site, and then calculate the means, standard errors, and their correlation of the incremental costs and effectiveness outcome across sites. The ICER CIs can be easily calculated by inputting these parameter estimates using the software. In addition, the Monte Carlo interval is particularly useful when the data from the original sample are not available and, thus, the bootstrap method is not feasible (e.g., in the systematic review).

The second advantage of the Monte Carlo interval is noteworthy. It provides the empirical distribution of the ICER, akin to the bootstrap method. Moreover, it offers confidence intervals in most situations, unlike Fieller's method. Researchers can easily identify negative and undefined ICERs in the empirical distribution. This is a significant benefit over Fieller's method, which struggles to provide meaningful confidence intervals when the treatment effect lacks statistical significance. The third advantage is that the Monte Carlo interval can be directly applied for robustness analysis by changing the input parameters (e.g., the cost-effectiveness correlation). These advantages make the Monte Carlo interval a powerful method that is feasible for most scenarios in practice, and useful for sensitivity analysis.

One limitation of the present study is that although we evaluated the Monte Carlo interval against some commonly used bootstrap methods (percentile, bias-corrected, and bias-corrected and accelerated), we did not test other methods, e.g., the Bayesian bootstrap (Rubin, 1981). Future research may investigate constructing credible intervals for the ICER using the Bayesian bootstrap method. Nevertheless, the percentile-based Monte Carlo interval has demonstrated good performance in constructing confidence intervals of the ICER.

Note that one limitation of applying ICER and ICER CIs for policy making is that the ICER assumes a linear relationship between the incremental costs and effectiveness outcomes while the incremental effectiveness results may not be scalable (i.e., an additional standard deviation increase in the effectiveness outcome may not be purchased by the same amount of money). For example, there may be a ceiling effect for impacts of the program or there may be a curvilinear relationship between the incremental costs and effectiveness outcomes. Questions about the value of ICER CIs to decision makers have been raised in medical literature and some researchers suggest that confidence surfaces are better suited for this decision making than CIs (e.g., Briggs & Fenn, 1998). One direction for future research is to investigate what measures or statistics (e.g., net monetary benefit) of cost-effectiveness analysis are more feasible and better than ICER CIs in education decision-making.

Furthermore, to our knowledge, there is no statistical method available for designing randomized cost-effectiveness studies with adequate statistical power based on ICER. Another direction for future research is to use the Monte Carlo interval for power analysis of the ICER based on two- or one-sided testing like the analysis of power for mediation effects using Monte Carlo intervals (e.g., Kelcey et al. 2017; 2020).

Finally, we echo the suggestion that “CEA would ideally include confidence intervals for the incremental cost estimate(s) and the resulting cost-effectiveness ratios” … “when an adequate number of cost estimates is available” in the *Standards for the Economic Evaluation of Educational and Social Programs* (Cost Analysis Standards Project, 2021, p.43). In addition, future empirical CEA studies are encouraged to report the variance (or standard error) of the incremental costs and the correlation of the incremental costs and incremental effectiveness when possible.

**References:**

Anderson, J.P., Bush, J.W., Chen, M., & Dolenc, D. (1986). Policy space areas and properties of benefit cost/utility analysis. *Journal of the American Medical Association, 255*(6): 794–795. 9.

Bai, F., Kelcey, B., Xie, Y., & Cox, K. (2023). Design and Analysis of Clustered Regression Discontinuity Designs for Probing Mediation Effects. *The Journal of Experimental Education,* 1–31. https://doi.org/10.1080/00220973.2023.2287445

Barrett, C. A., Truckenmiller, A. J., & Eckert, T. L. (2020). Performance feedback during writing instruction: A cost-effectiveness analysis. *School Psychology, 35*(3), 193–200. https://doi.org/10.1037/spq0000356

Barrett, C. A., & VanDerHeyden, A. M. (2020). A cost-effectiveness analysis of classwide math intervention. *Journal of school psychology, 80*, 54–65. https://doi.org/10.1016/j.jsp.2020.04.002

Black, W.C. (1990). The CE plane: a graphic representation of cost-effectiveness. *Medical Decision Making*, 10: 212–214.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2018). *Cost-benefit analysis: Concepts and practice* (Fifth edition). Cambridge University Press, Cambridge, United Kingdom ; New York, NY.

Borman, G. D., & Hewes, G. M. (2002). The Long-Term Effects and Cost-Effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(4), 243–266. https://doi.org/10.3102/01623737024004243

Bowden, A. B., & 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

Bowden, A. B., Shand, R., Belfield, C. R., Wang, A., & Levin, H. M. (2017). Evaluating Educational Interventions That Induce Service Receipt: A Case Study Application of City Connects. *American Journal of Evaluation, 38*(3), 405–419. https://doi.org/10.1177/1098214016664983

Briggs, A., & Fenn, P. (1998). Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics, 7*(8), 723-740.

Briggs, A. H., O'Brien, B. J., & Blackhouse, G. (2002). Thinking outside the box: recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual review of public health, 23*, 377–401. https://doi.org/10.1146/annurev.publhealth.23.100901.140534

Briggs, A. H., Mooney, C. Z., & Wonderling, D. E. (1999). Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in medicine*, *18*(23), 3245–3262. https://doi.org/10.1002/(sici)1097-0258(19991215)18:23<3245::aid-sim314>3.0.co;2-2

Chaudhary, M. A., & Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in medicine, 15*(13), 1447–1458. https://doi.org/10.1002/(SICI)1097-0258(19960715)15:13<1447::AID-SIM267>3.0.CO;2-V

Clarke, B., Cil, G., Smolkowski, K., Sutherland, M., Turtura, J., Doabler, C. T., Fien, H. & Baker, S. K. (2020) Conducting a Cost-Effectiveness Analysis of an Early Numeracy

Intervention. *School Psychology Review, 49* (4), 359-373, DOI: 10.1080/2372966X.2020.1761236

Cil, G., Chaparro, E. A., Dennis, C., & Smolkowski, K. (2023). The cost-effectiveness of an English language curriculum for middle school English learners. *School psychology, 38*(1), 48–58. https://doi.org/10.1037/spq0000515

Cost Analysis Standards Project. (2021). *Standards for the economic evaluation of educational and social programs*. American Institutes for Research. Retrieved April 3rd, 2022 from https://www.air.org/sites/default/files/Standards-for-the-Economic-Evaluation-of-Educational-and-Social-Programs-CASP-May-2021.pdf

Cox, K. & Kelcey, B. (2023). A Partial Posterior p value Test for Multilevel Mediation. *Statistica Neerlandica, 77, 408-428.*

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7*(1), 1–26. DOI: 10.1214/aos/1176344552

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other resampling methods. *Biometrika, 68*, 589–599.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Philadelphia, PA: Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association, 82*(397), 171-185. doi: 10.1080/01621459.1987.10478410

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fieller, E. C. (1954). Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological), 16*(2), 175–185. http://www.jstor.org/stable/2984043

Finster, M., Decker-Woodrow, L., Booker, B., Mason, C. A., Tu, S. & Lee, J. (2023). Cost-Effectiveness of Algebraic Technological Applications, *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2023.2269918

Gray, A. M., Sirinides, P. M., Fink, R. E., & Bowden, A. B. (2021). Integrating literacy and science instruction in kindergarten: Results from the efficacy study of Zoology One. *Journal of Research on Educational Effectiveness*, 1–27. https://doi.org/10.1080/19345747.2021.1938313

Guryan, J., Christenson, S., Cureton, A., Lai, I., Ludwig, J., Schwarz, C., Shirey, E., & Turner, M.C. (2021), The Effect of Mentoring on School Attendance and Academic Outcomes: A Randomized Evaluation of the Check & Connect Program. *Journal of Policy Analysis and Management, 40* (3), 841-882. https://doi.org/10.1002/pam.22264

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-Effectiveness Analysis of Early Reading Programs: A Demonstration With Recommendations for Future Research, *Journal of Research on Educational Effectiveness, 9*(1), 30-53. doi: 10.1080/19345747.2015.1055639

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

Hunter, L. J., DiPerna, J. C., Hart, S. C., & Crowley, M. (2018). At what cost? Examining the cost effectiveness of a universal social-emotional learning program. *School psychology Quarterly, 33*(1), 147–154. https://doi.org/10.1037/spq0000232

Institute of Education Sciences (IES). (2020). *Cost Analysis: A Toolkit* (IES 2020-001). U.S. Department of Education. Washington, DC: Institute of Education Sciences. Retrieved April 3rd, 2022 from https://ies.ed.gov/seer/pdf/IES_Cost_Analysis_Starter_Kit_V1.pdf.

Institute of Education Sciences (IES). (2024). *Request for applications: Education research grants program*. U.S. Department of Education. https://ies.ed.gov/funding/pdf/2024_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

Jiang, G., Wu, J. & Williams, G.R. (2000). Fieller's Interval and the Bootstrap-Fieller Interval for the Incremental Cost-Effectiveness Ratio. *Health Services & Outcomes Research Methodology 1*, 291–303. https://doi.org/10.1023/A:1011499328061

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical power for causally-defined individual and contextual indirect effects in group-randomized Trials. *Journal of Educational and Behavioral Statistics, 24* (5), 499-530. doi: 10.3102/1076998617695506

Kelcey, B., Spybrook, J., Dong, N., & Bai, F. (2020). Cross-level mediation in school-randomized studies of teacher development: Experimental design and power. *Journal of Research on Educational Effectiveness, 13* (3), 459-487. doi: 10.1080/19345747.2020.1726540

Kim, J. S., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., Collins, P. & Land, R. E. (2011) A Randomized Experiment of a Cognitive Strategies Approach to Text-Based Analytical Writing for Mainstreamed Latino English Language Learners in Grades

6 to 12. *Journal of Research on Educational Effectiveness, 4* (3), 231-263, DOI: 10.1080/19345747.2010.523513

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

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., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2018). *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. doi: 10.3102/1076998620911916

Li, W., Dong, N., Maynard, R. A., Spybrook, J., & Kelcey, B. (2023). Experimental design and statistical power for Cluster Randomized Cost-Effectiveness Trials. *Journal of Research on Educational Effectiveness, 16*(4), 681-706, DOI: 10.1080/19345747.2022.2142177

Mullahy, J., & Manning, W. (1995). Statistical issues in cost–effectiveness analyses. In F. Sloan (Ed.), *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies* (pp. 149-184). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511625817.008

O'Brien, B. J., Drummond, M. F., Labelle, R. J., & Willan, A. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical care, 32*(2), 150–163. https://doi.org/10.1097/00005650-199402000-00006

Polsky, D., Glick, H. A., Willke, R., & Schulman, K. (1997). Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health economics, 6*(3), 243–252. https://doi.org/10.1002/(sici)1099-1050(199705)6:3<243::aid-hec269>3.0.co;2-z

Preacher, K. J. & Selig, J. P. (2012) Advantages of Monte Carlo Confidence Intervals for Indirect Effects, *Communication Methods and Measures, 6*(2), 77-98, DOI: 10.1080/19312458.2012.679848

Rubin, R. D. (1981). The Bayesian Bootstrap. *The Annals of Statistics, 9*(1), 130-134. DOI: 10.1214/aos/1176345338

SAS Institute Inc. (2007). *Jackknife and Bootstrap Analyses*. Available at https://support.sas.com/kb/24/982.html

Scammacca, N., Swanson, E., Vaughn, S. & Roberts, G. (2020) Cost-Effectiveness of a Grade 8 Intensive Reading and Content Learning Intervention. *School Psychology Review, 49*(4), 374-385, DOI: 10.1080/2372966X.2020.1760691

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, *80*, 360–361.

Stinnett, A. A. & Mullahy, J. (1998). *Net Health Benefits: A New Framework for the Analysis of Uncertainty in Cost-Effectiveness Analysis*. NBER Working Paper No. t0227, Available at SSRN: https://ssrn.com/abstract=226637

Unlu, F., Edmunds, J., Fesler, L., & Glennie, B. (2015). A preliminary assessment of the cost and benefit of the North Carolina's Early College High School Model and its impact on postsecondary enrollment and earned college credit. [Paper presentation]. Spring 2015 Conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC, United States.

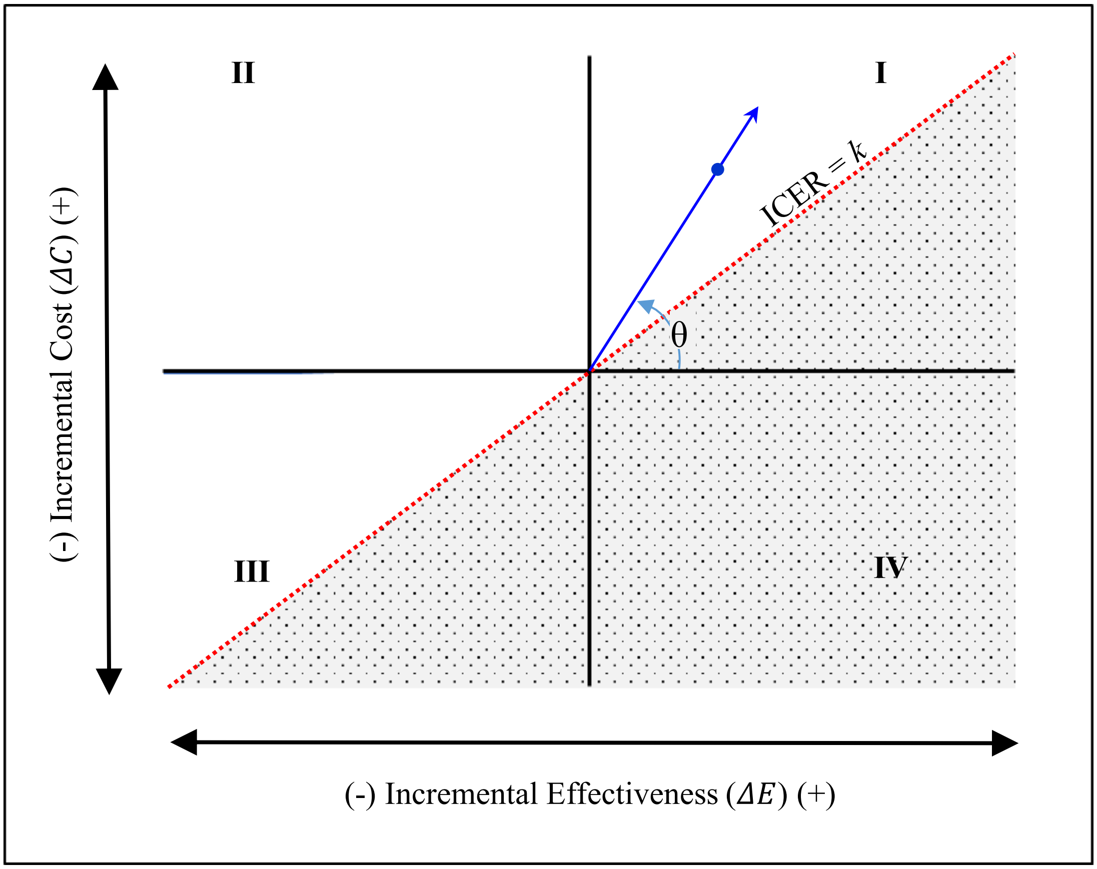
van Hout, B. A., Al, M. J., Gordon, G. S., & Rutten, F. F. (1994). Costs, effects and C/E-ratios alongside a clinical trial. *Health economics, 3*(5), 309–319. https://doi.org/10.1002/hec.4730030505

Wakker, P. & Klaassen, M.P. (1995). Confidence intervals for cost/effectiveness ratios. *Health Economics, 4*(5): 373-381. https://doi.org/10.1002/hec.4730040503

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

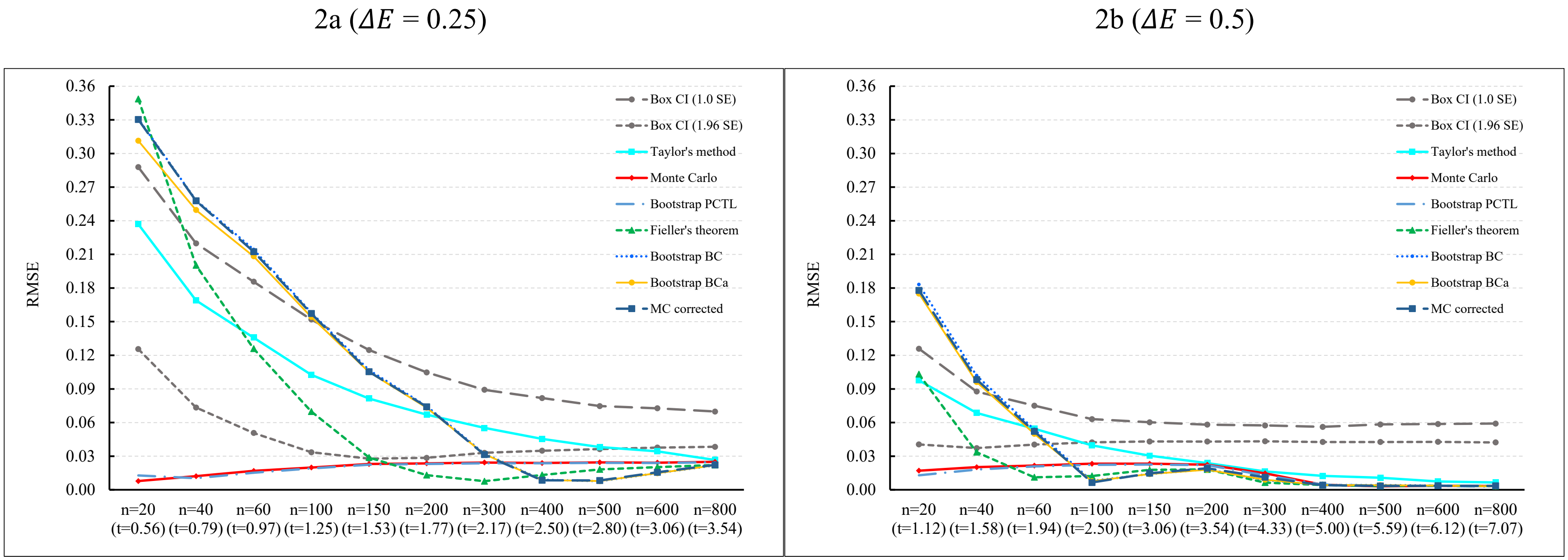
Figures and Tables

Figure 1: The $\Delta E - \Delta C$ Plane


II
I
ICER = k
θ
III
IV
(-) Incremental Cost ($\Delta C$) (+)
(-) Incremental Effectiveness ($\Delta E$) (+)

Note: The figure has been used in literature (e.g., Anderson et al., 1986; Black, 1990).

Figure 2: RMSE of 95% confidence interval coverage as a function of sample size

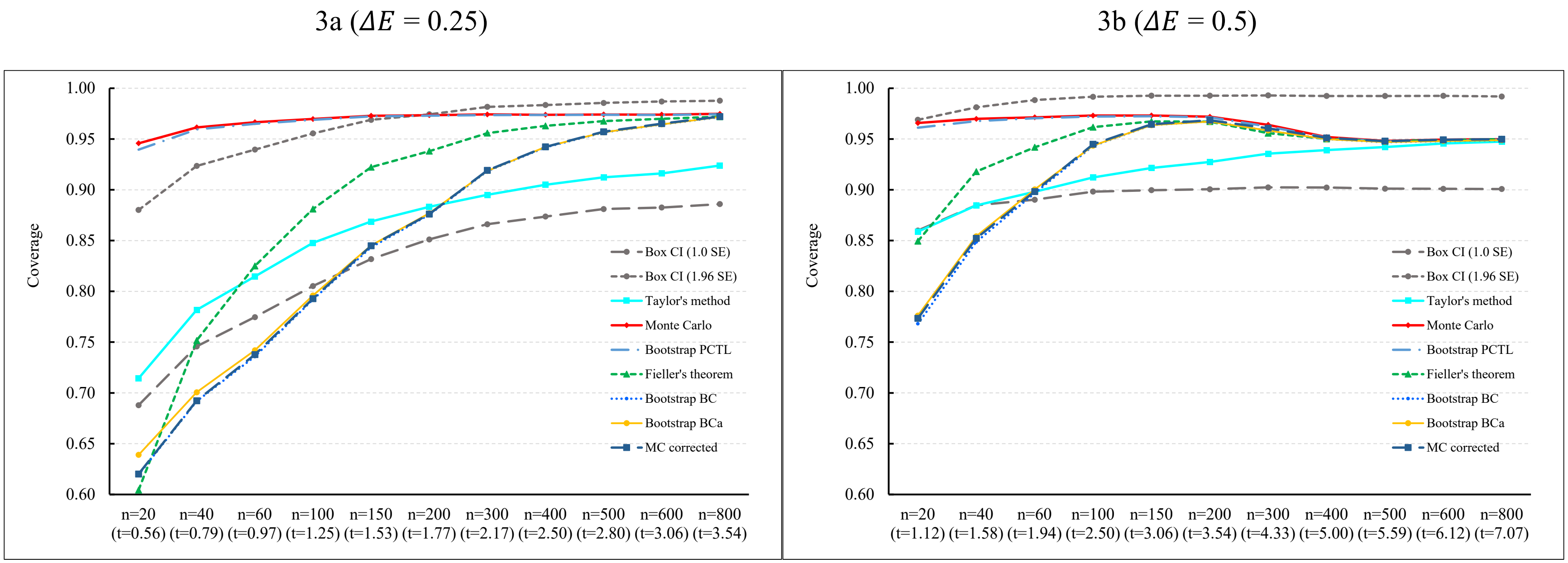


2a ($\Delta E$ = 0.25)

2b ($\Delta E$ = 0.5)

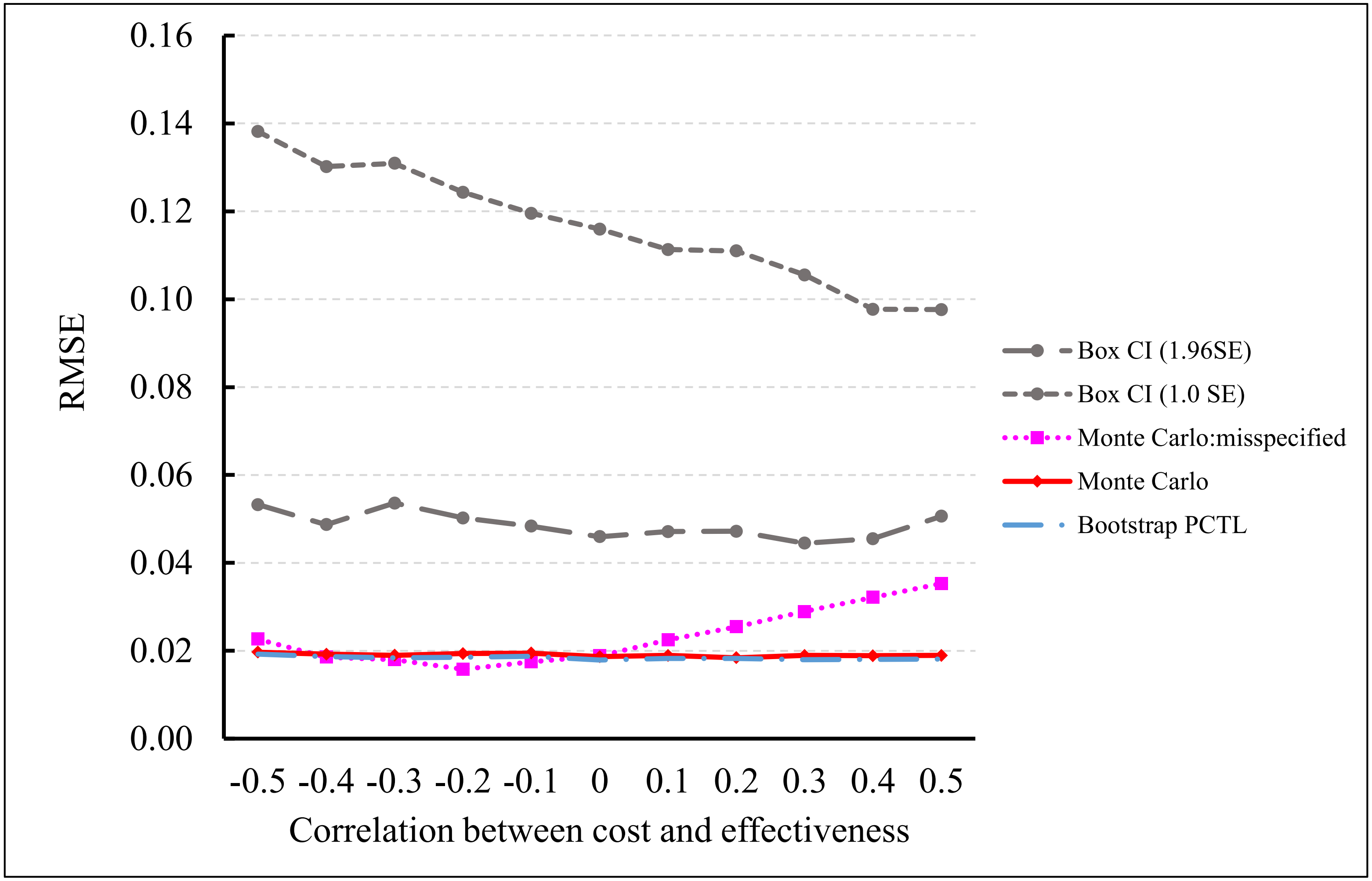
Note: RMSE was calculated across 22 scenarios (2 cost distributions × 11 correlations). “t” refers to the *t*-statistic of the treatment effect size for the effectiveness measure.

Figure 3: 95% confidence interval coverage as a function of sample size



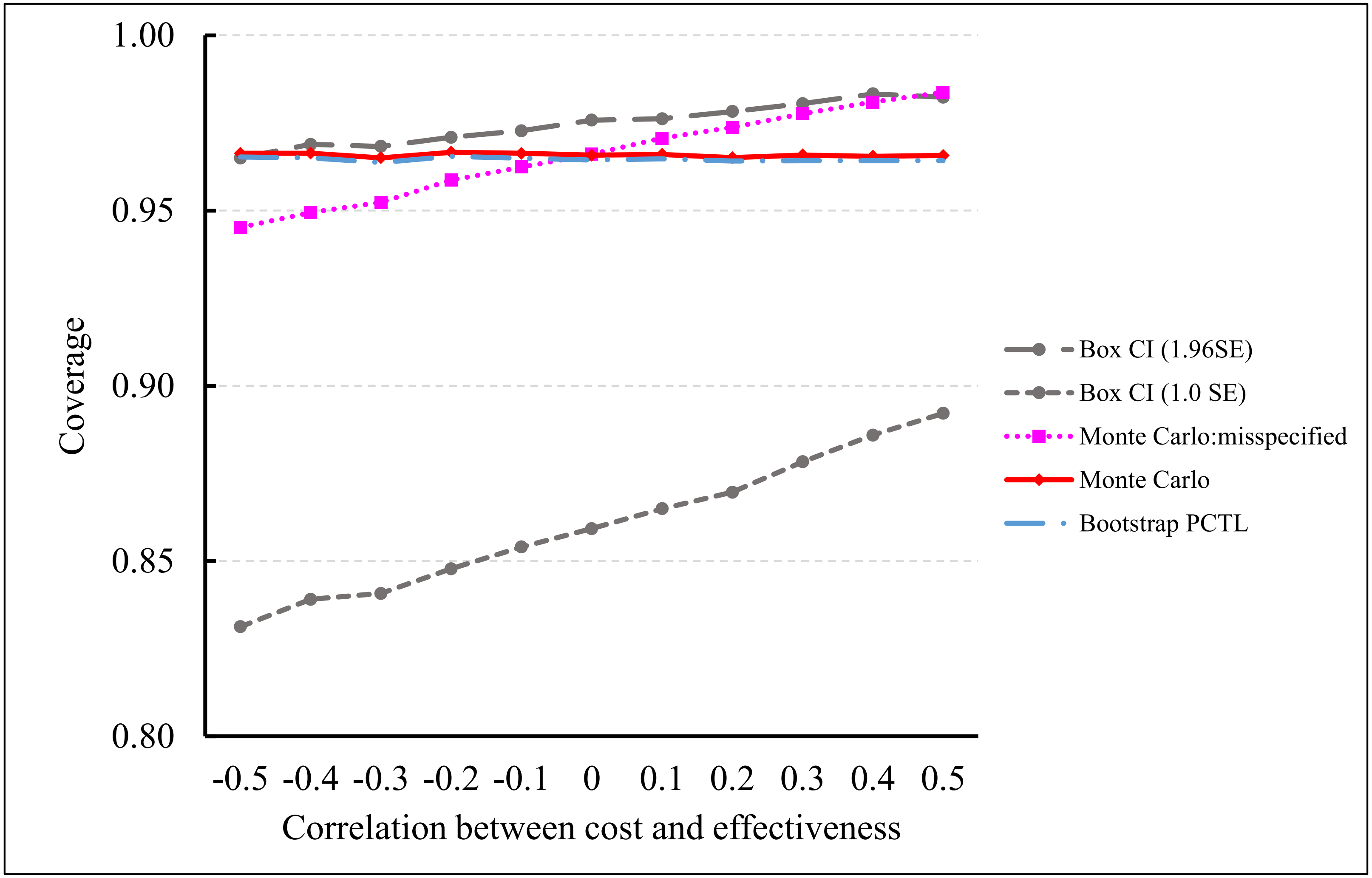
Note: Coverage rate was calculated across 22 scenarios (2 cost distributions × 11 correlations). “t” refers to the *t*-statistic of the treatment effect size for the effectiveness measure.

Figure 4: RMSE of 95% confidence interval coverage as a function of cost-effectiveness correlation



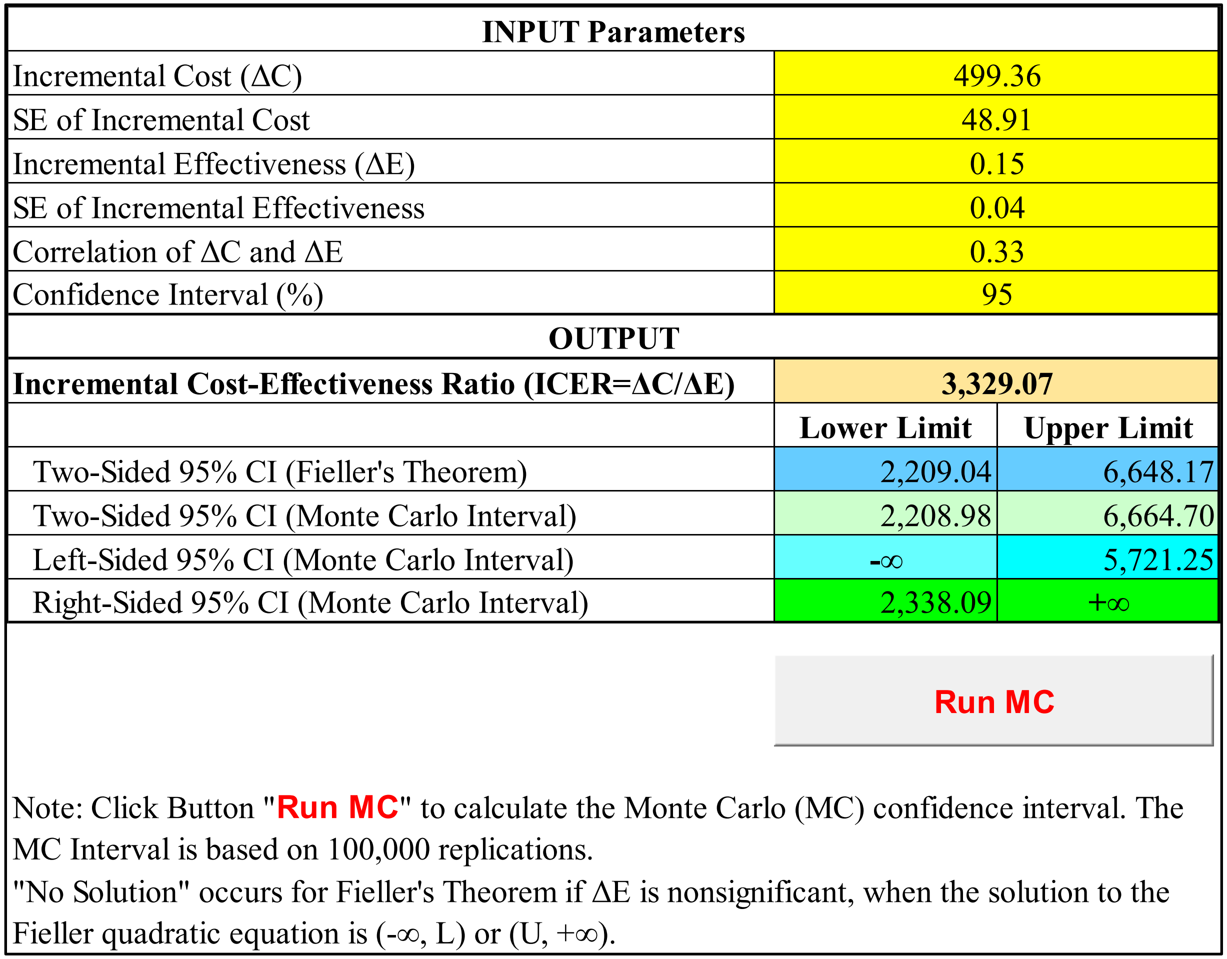
Note: RMSE was calculated across 44 scenarios (11 sample sizes × 2 cost distributions × 2 $\Delta E$).

Figure 5: 95% confidence interval coverage as a function of cost-effectiveness correlation



Note: RMSE was calculated across 44 scenarios (11 sample sizes × 2 cost distributions × 2 $\Delta E$).

Figure 6: Screenshot of the ICER CI Calculator



| INPUT Parameters | | |
|---|---|---|
| Incremental Cost (ΔC) | 499.36 | |
| SE of Incremental Cost | 48.91 | |
| Incremental Effectiveness (ΔE) | 0.15 | |
| SE of Incremental Effectiveness | 0.04 | |
| Correlation of ΔC and ΔE | 0.33 | |
| Confidence Interval (%) | 95 | |
| **OUTPUT** | | |
| **Incremental Cost-Effectiveness Ratio (ICER=ΔC/ΔE)** | **3,329.07** | |
| | **Lower Limit** | **Upper Limit** |
| Two-Sided 95% CI (Fieller's Theorem) | 2,209.04 | 6,648.17 |
| Two-Sided 95% CI (Monte Carlo Interval) | 2,208.98 | 6,664.70 |
| Left-Sided 95% CI (Monte Carlo Interval) | -∞ | 5,721.25 |
| Right-Sided 95% CI (Monte Carlo Interval) | 2,338.09 | +∞ |

Note: Click Button "**Run MC**" to calculate the Monte Carlo (MC) confidence interval. The MC Interval is based on 100,000 replications.
"No Solution" occurs for Fieller's Theorem if ΔE is nonsignificant, when the solution to the Fieller quadratic equation is (-∞, L) or (U, +∞).

Table 1: The summary results of 484 simulation scenarios

| Method | | RMSE | Coverage rate | Width | Symmetry |
|---|---|---|---|---|---|
| Box CI | 1.96 SE | 0.049 | 0.975<br>(0.715, 1.000) | 309118<br>(15081, 9M) | 0.00<br>(0.00, 0.00) |
| | 1.0 SE | 0.117 | 0.860<br>(0.523, 0.975) | 219683<br>(12225, 30M) | 0.00<br>(0.00, 0.00) |
| Taylor's method | | 0.084 | 0.890<br>(0.680, 0.96) | 149B<br>(5386, 63T) | 0.00<br>(0.00, 0.09) |
| Bootstrap | Bias-corrected | 0.119 | 0.884<br>(0.583, 0.980) | 21M<br>(5828, 1.4B) | 0.38<br>(0.04, 0.59) |
| | Bias-corrected & accelerated | 0.114 | 0.887<br>(0.603, 0.98) | 18M<br>(5827, 841M) | 0.37<br>(0.04, 0.58) |
| | Percentile | 0.018 | 0.965<br>(0.924, 0.983) | 148462<br>(5829, 453887) | 0.11<br>(0, 0.56) |
| Fieller's theorem | | 0.095 | 0.914<br>(0.530, 0.980) | 251927<br>(5825, 15M) | 0.12<br>(0.00, 0.57) |
| Monte Carlo interval | Percentile | 0.019 | 0.966<br>(0.934, 0.983) | 147065<br>(5825, 450598) | 0.11<br>(0, 0.57) |
| | Bias-corrected | 0.118 | 0.885<br>(0.585, 0.981) | 460M<br>(5825, 26B) | 0.38<br>(0.04, 0.60) |
| | Misspecified correlation=0 | 0.024 | 0.966<br>(0.888, 0.997) | 174296<br>(6497, 415561) | 0.10<br>(0.00, 0.62) |

*N* = 484. Within the parathesis are the minimal and maximum values. “M” refers to million, “B” refers to billion, and “T” refers to trillion.

Table 2: The results of regressing the absolute value of the bias of the 95% coverage rate on the simulation factors

| Variables | Box CI (1.96 SE) | Box CI (1.0 SE) | Taylor's method | Bootstrap bias-corrected | Bootstrap bias-corrected & accelerated | Bootstrap percentile | Fieller's theorem | Monte Carlo (percentile) | Monte Carlo (bias-corrected) | Monte Carlo: misspecified correlation=0 | Monte Carlo: misspecified correlation=0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | **0.041*** | **0.091*** | **0.061*** | **0.074*** | **0.071*** | **0.016*** | **0.048*** | **0.017*** | **0.073*** | **0.021*** | **0.021*** |
| Sample Size (1,000) | **-0.013*** | **-0.115*** | **-0.144*** | **-0.242*** | **-0.231*** | **-0.006*** | **-0.155*** | **-0.007*** | **-0.239*** | **0.005** | **0.005*** |
| Effectiveness Effect Size | -0.008 | **-0.263*** | **-0.234*** | **-0.302*** | **-0.293*** | **-0.029*** | **-0.237*** | **-0.026*** | **-0.303*** | 0.002 | 0.002 |
| Cost-Effect Correlation | 0.004 | **-0.053*** | -0.007 | -0.004 | -0.004 | -0.001 | -0.005 | <0.001 | -0.005 | **0.019*** | **0.019*** |
| Lognormal Cost | 0.003 | **-0.071*** | **-0.015*** | -0.004 | -0.004 | -0.001 | -0.005 | -0.001 | -0.005 | **0.004*** | **0.004*** |
| Squared Cost-Effect Correlation | na | na | na | na | na | na | na | na | na | na | **0.031*** |
| $R^2$ | 0.02 | 0.63 | 0.65 | 0.56 | 0.57 | 0.21 | 0.35 | 0.18 | 0.56 | 0.31 | 0.37 |

Note: $N = 484$. *$p < 0.01$.
All predictors are mean centered. The intercept represents the mean absolute value of the bias.

Table 3: Width by the $\Delta E$ and sample size

| | $\Delta E$ = 0.25 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | n=20<br>(t=0.56) | n=40<br>(t=0.79) | n=60<br>(t=0.97) | n=100<br>(t=1.25) | n=150<br>(t=1.53) | n=200<br>(t=1.77) | n=300<br>(t=2.17) | n=400<br>(t=2.50) | n=500<br>(t=2.80) | n=600<br>(t=3.06) | n=800<br>(t=3.54) |
| Box CI (1.96 SE) | 263,684 | 413,344 | 870,427 | 385,835 | 692,307 | 575,276 | 661,146 | 374,906 | 244,068 | 475,753 | 208,643 |
| Box CI (1.0 SE) | 175,182 | 405,816 | 342,171 | 297,378 | 745,588 | 266,322 | 332,947 | 1M | 84,076 | 74,256 | 45,843 |
| Taylor's method | 3T | 4B | 13B | 26B | 5B | 1B | 94M | 156B | 13M | 24M | 134,457 |
| Bootstrap bias-corrected | 31M | 50M | 124M | 86M | 31M | 38M | 19M | 10M | 7M | 4M | 1M |
| Bootstrap bias-corrected & accelerated | 25M | 39M | 90M | 83M | 29M | 37M | 18M | 10M | 7M | 4M | 1M |
| Bootstrap percentile | 255,237 | 316,643 | 351,621 | 383,762 | 380,935 | 364,689 | 301,303 | 238,308 | 184,696 | 140,480 | 86,402 |
| Fieller's theorem | 269,538 | 267,743 | 278,766 | 295,430 | 327,436 | 545,315 | 668,887 | 241,072 | 389,886 | 898,588 | 122,261 |
| Monte Carlo (percentile) | 250,758 | 312,152 | 347,380 | 379,566 | 377,709 | 360,916 | 298,656 | 235,875 | 182,920 | 139,558 | 85,544 |
| Monte Carlo (bias-corrected) | 1B | 1B | 1B | 1B | 975M | 579M | 2B | 240M | 332M | 132M | 8M |
| Monte Carlo: misspecified correlation=0 | 252,669 | 313,074 | 347,874 | 379,761 | 377,750 | 361,092 | 298,764 | 235,947 | 182,987 | 139,620 | 85,601 |

| | $\Delta E$ = 0.50 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | n=20<br>(t=1.12) | n=40<br>(t=1.58) | n=60<br>(t=1.94) | n=100<br>(t=2.50) | n=150<br>(t=3.06) | n=200<br>(t=3.54) | n=300<br>(t=4.33) | n=400<br>(t=5.00) | n=500<br>(t=5.59) | n=600<br>(t=6.12) | n=800<br>(t=7.07) |
| Box CI (1.96 SE) | 365,161 | 406,564 | 294,469 | 209,281 | 144,163 | 95,829 | 39,866 | 24,378 | 20,692 | 18,358 | 16,455 |
| Box CI (1.0 SE) | 143,376 | 118,552 | 101,322 | 68,045 | 44,366 | 38,958 | 16,733 | 14,955 | 14,025 | 13,498 | 12,841 |
| Taylor's method | 195B | 62M | 42M | 8M | 84M | 133,505 | 12,824 | 10,204 | 8,838 | 7,917 | 6,704 |
| Bootstrap bias-corrected | 23M | 18M | 22M | 6M | 1M | 601,253 | 42,626 | 13,774 | 10,591 | 9,006 | 7,315 |
| Bootstrap bias-corrected & accelerated | 19M | 12M | 19M | 4M | 1M | 585,903 | 41,252 | 13,748 | 10,581 | 9,002 | 7,314 |
| Bootstrap percentile | 200,748 | 196,175 | 173,001 | 119,573 | 70,772 | 43,622 | 20,550 | 13,429 | 10,558 | 8,999 | 7,314 |
| Fieller's theorem | 147,715 | 186,175 | 494,327 | 141,377 | 140,765 | 59,983 | 25,853 | 14,126 | 10,821 | 9,028 | 7,309 |
| Monte Carlo (percentile) | 198,367 | 194,549 | 171,690 | 118,874 | 70,495 | 43,566 | 20,566 | 13,401 | 10,566 | 9,016 | 7,309 |
| Monte Carlo (bias-corrected) | 469M | 424M | 336M | 105M | 40M | 4M | 68,782 | 13,747 | 10,597 | 9,027 | 7,309 |
| Monte Carlo: misspecified correlation=0 | 199,175 | 194,448 | 171,578 | 118,837 | 70,534 | 43,646 | 20,633 | 13,467 | 10,623 | 9,071 | 7,354 |

Note: Width was calculated across 22 scenarios (2 cost distributions × 11 correlations). “M” refers to million, “B” refers to billion, and “T” refers to trillion. “t” refers to the *t*-statistic of the treatment effect size for the effectiveness measure.

Table 4: Symmetry by the $\Delta E$ and sample size

| | $\Delta E$ = 0.25 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | n=20 (t=0.56) | n=40 (t=0.79) | n=60 (t=0.97) | n=100 (t=1.25) | n=150 (t=1.53) | n=200 (t=1.77) | n=300 (t=2.17) | n=400 (t=2.50) | n=500 (t=2.80) | n=600 (t=3.06) | n=800 (t=3.54) |
| Box CI (1.96 SE) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Box CI (1.0 SE) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Taylor's method | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Bootstrap bias-corrected | 0.20 | 0.28 | 0.34 | 0.42 | 0.48 | 0.51 | 0.51 | 0.44 | 0.34 | 0.23 | 0.10 |
| Bootstrap bias-corrected & accelerated | 0.21 | 0.28 | 0.33 | 0.41 | 0.47 | 0.51 | 0.50 | 0.44 | 0.34 | 0.22 | 0.10 |
| Bootstrap percentile | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Fieller's theorem | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| Monte Carlo (percentile) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Monte Carlo (bias-corrected) | 0.20 | 0.28 | 0.34 | 0.42 | 0.48 | 0.52 | 0.51 | 0.45 | 0.34 | 0.22 | 0.09 |
| Monte Carlo: misspecified correlation=0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | $\Delta E$ = 0.50 | | | | | | | | | | |
| Method | n=20 (t=1.12) | n=40 (t=1.58) | n=60 (t=1.94) | n=100 (t=2.50) | n=150 (t=3.06) | n=200 (t=3.54) | n=300 (t=4.33) | n=400 (t=5.00) | n=500 (t=5.59) | n=600 (t=6.12) | n=800 (t=7.07) |
| Box CI (1.96 SE) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Box CI (1.0 SE) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Taylor's method | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| Bootstrap bias-corrected | 0.37 | 0.46 | 0.49 | 0.41 | 0.23 | 0.17 | 0.37 | 0.48 | 0.50 | 0.51 | 0.48 |
| Bootstrap bias-corrected & accelerated | 0.38 | 0.44 | 0.46 | 0.39 | 0.22 | 0.16 | 0.36 | 0.47 | 0.50 | 0.51 | 0.48 |
| Bootstrap percentile | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.08 | 0.29 | 0.46 | 0.50 | 0.51 | 0.48 |
| Fieller's theorem | 0.00 | 0.00 | 0.00 | 0.01 | 0.06 | 0.14 | 0.41 | 0.49 | 0.50 | 0.51 | 0.48 |
| Monte Carlo (percentile) | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.07 | 0.28 | 0.47 | 0.50 | 0.51 | 0.48 |
| Monte Carlo (bias-corrected) | 0.38 | 0.47 | 0.49 | 0.42 | 0.24 | 0.17 | 0.35 | 0.48 | 0.50 | 0.51 | 0.48 |
| Monte Carlo: misspecified correlation=0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.27 | 0.46 | 0.49 | 0.50 | 0.48 |

Note: Symmetry was calculated across 22 scenarios (2 cost distributions × 11 correlations). “t” refers to the $t$-statistic of the treatment effect size for the effectiveness measure.

## Supplemental Material:

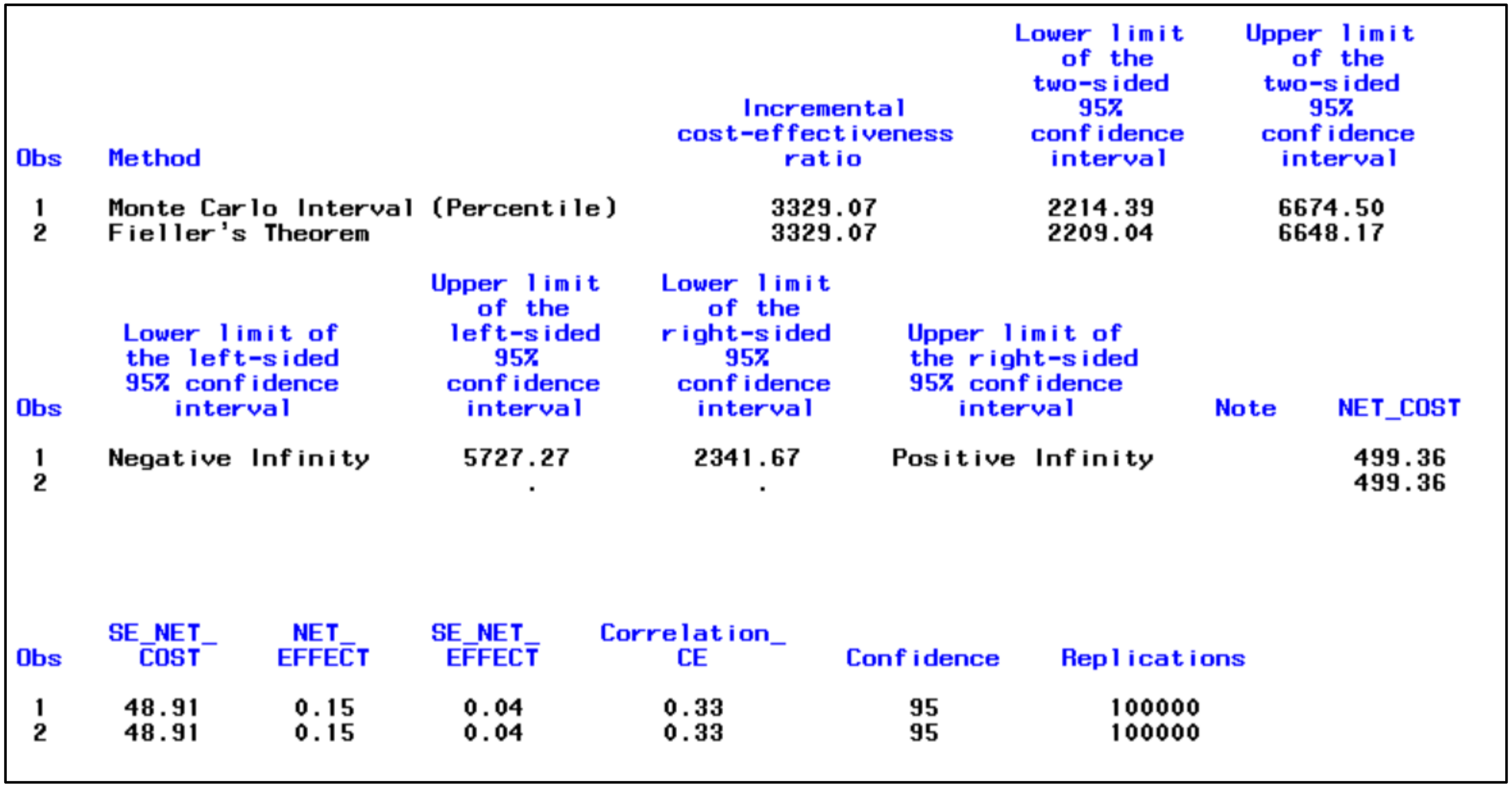
### SAS Macro for Calculating ICER CIs

```
%MACRO
ICER_CIs_Calculator(NET_COST,SE_NET_COST,NET_EFFECT,SE_NET_EFFECT,Correlation
_CE,CONFID, REPLICATIONS);

/*** Monte Carlo Interval ***/
/* Create net cost-effectiveness variance-covariance matrix, and simulated
data for Monte Carlo Interval*/
%Let cost_var = %sysevalf(&SE_NET_COST*&SE_NET_COST);
/* Variance of net cost estimate */
%Let effect_var = %sysevalf(&SE_NET_EFFECT*&SE_NET_EFFECT);
/* Variance of net effect estimate */
%Let ce_cov = %sysevalf(&Correlation_CE*&SE_NET_COST*&SE_NET_EFFECT);
/* Covariance */
%Let ICER = %sysevalf(&NET_COST/&NET_EFFECT);
%Let pctl = %sysevalf((100-&CONFID)/2);
        /* Two-sided CI: left limit percentile */
%Let pctu = %sysevalf(&CONFID+(100-&CONFID)/2);
        /* Two-sided CI: right limit percentile */
%Let pctu_L_sided = %sysevalf(&CONFID);
        /* Left-sided CI: right limit percentile */
%Let pctl_R_sided = %sysevalf(100-&CONFID);
        /* Right-sided CI: left limit percentile */

proc iml;
mu = { &NET_COST, &NET_EFFECT };
sigma= { &cost_var &ce_cov,
&ce_cov &effect_var};
call vnormal(ce, mu, sigma, &REPLICATIONS);
create mcdata (RENAME=(COL1=cost COL2=effect)) from ce ;
append from ce;
quit;

data mcdata;
set mcdata;
Lable ICER_sim = "Incremental cost-effectiveness ratios from simulation";
ICER_sim = cost/effect;
run;

* Get percentile confidence limits for simulated data ;
proc univariate data=mcdata noprint;
      var ICER_sim;
      output out=ci_mcdata pctlpts= &pctl &pctu &pctl_R_sided &pctu_L_sided
pctlpre=a pctlname=lcl ucl lcl_R_sided ucl_L_sided n=n;
run;

data ci_mcdata;
alcl_L_sided = 'Negative Infinity';
aucl_R_sided = 'Positive Infinity';
retain alcl aucl alcl_L_sided aucl_L_sided alcl_R_sided aucl_R_sided;
```

```
drop n;
length Method $50;
Method="Monte Carlo Interval (Percentile)";
set ci_mcdata;
run;

/*** Fieller's theorem ***/

data Fieller;
drop b2 a c;
length method $50;

b2= &NET_COST*&NET_EFFECT-
quantile('NORMAL',&pctu/100)**2*&Correlation_CE*&SE_NET_COST*&SE_NET_EFFECT;
a = &NET_EFFECT*&NET_EFFECT-
quantile('NORMAL',&pctu/100)**2*&SE_NET_EFFECT*&SE_NET_EFFECT;
c = &NET_COST*&NET_COST-
quantile('NORMAL',&pctu/100)**2*&SE_NET_COST*&SE_NET_COST;

if a le 0 then do;
Note = "No Solution";
alcl = .;
aucl = .;
end;

if a > 0 then do;
alcl = (b2-sqrt(b2*b2-a*c))/a ;
aucl = (b2+sqrt(b2*b2-a*c))/a ;
end;

Method="Fieller's Theorem";
output;
run;

data ICER_CIs;
retain Method;
ICER = &ICER;
Lable ICER = "Incremental cost-effectiveness ratio";
Lable alcl = "Lower limit of the two-sided &CONFID% confidence interval";
Lable aucl = "Upper limit of the two-sided &CONFID% confidence interval";
Lable alcl_L_sided = "Lower limit of the left-sided &CONFID% confidence
interval";
Lable aucl_L_sided = "Upper limit of the left-sided &CONFID% confidence
interval";
Lable alcl_R_sided = "Lower limit of the right-sided &CONFID% confidence
interval";
Lable aucl_R_sided = "Upper limit of the right-sided &CONFID% confidence
interval";

set ci_mcdata Fieller;
NET_COST = &NET_COST;
SE_NET_COST = &SE_NET_COST;
NET_EFFECT = &NET_EFFECT;
SE_NET_EFFECT = &SE_NET_EFFECT;
Correlation_CE = &Correlation_CE;
Confidence=&CONFID;
Replications = &REPLICATIONS;
```

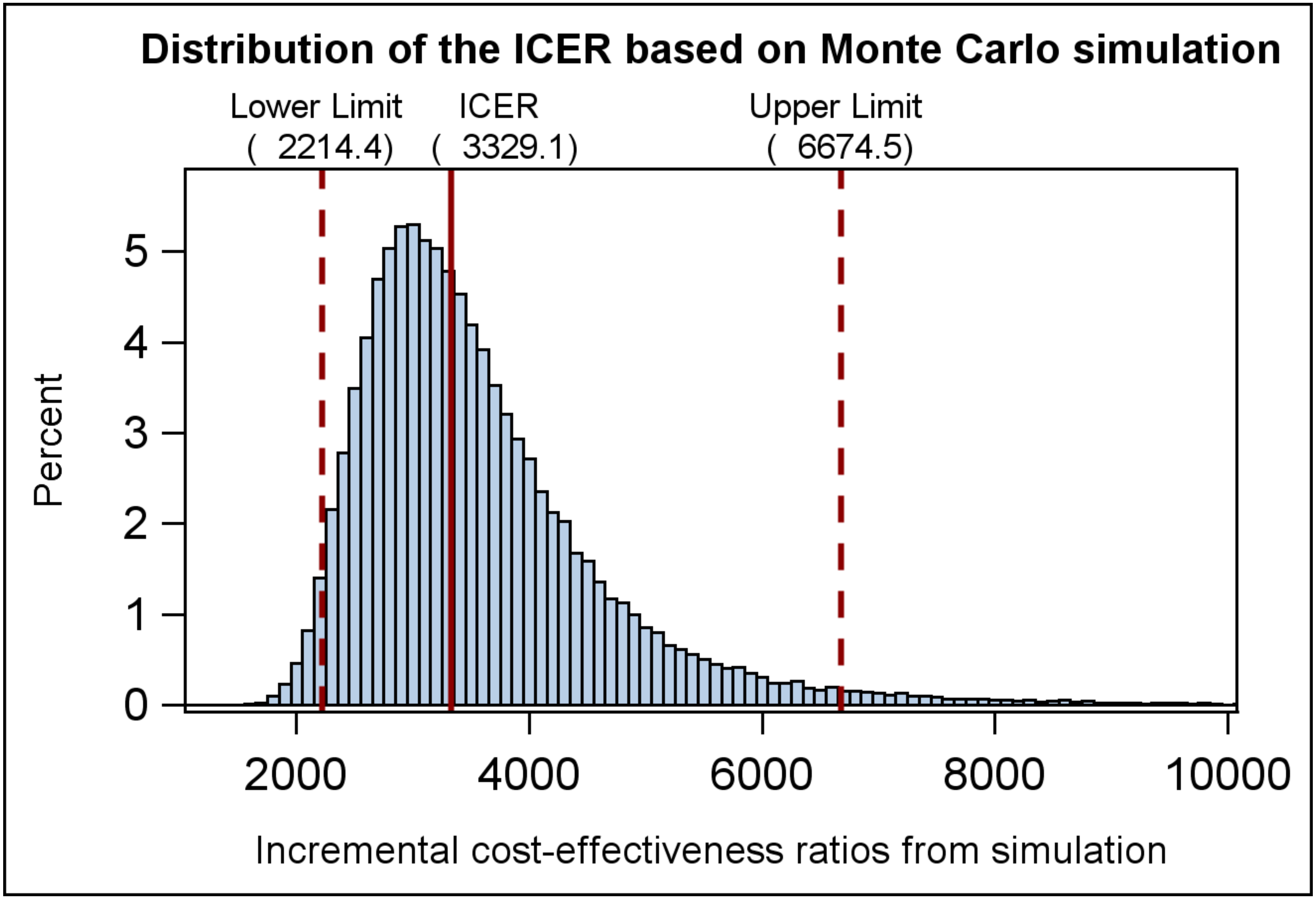
```
run;

Proc print data = ICER_CIs label;
run;

proc sql noprint;
select
round(ICER,.1),
round(alcl,.1),
round(aucl,.1),
alcl - (ICER - alcl) ,
aucl + (aucl - ICER) ,
10000
into
:ICER,
:alcl_PCTL,
:aucl_PCTL,
:min_xaxis ,
:max_xaxis ,
:n_bins
from Icer_cis where Method="Monte Carlo Interval (Percentile)";
quit;

ods listing image_dpi=300;
ods graphics / width=4.4 in height=3 in;

ods listing gpath='C:\ICER';

ods graphics / imagename="dist_ICER_sim" imagefmt=png;
title 'Distribution of the ICER based on Monte Carlo simulation' ;
proc sgplot data=mcdata ;
xaxis
  valueattrs=(color=black size=12pt) min = &min_xaxis max = &max_xaxis;
yaxis
  valueattrs=(color=black size=12pt) offsetmax=0.1;
   histogram ICER_sim  / scale=percent nbins = &n_bins;
    refline &alcl_PCTL &aucl_PCTL / axis=x lineattrs=(thickness=2
color=darkred pattern=dash)
            label=("Lower Limit /(&alcl_PCTL)"  "Upper Limit /(&aucl_PCTL)")
splitchar="/";
      refline &ICER  / axis=x lineattrs=(thickness=2 color=darkred )
                  label=("ICER /(&ICER)") splitchar="/";
run;
title;

%MEND;


%ICER_CIs_Calculator(499.36,48.91,0.15,0.04,0.33,95, 100000);
```

Figure S1: Screenshot of the output using the SAS Macro



```
                                                                 Lower limit    Upper limit
                                                                    of the         of the
                                                                  two-sided      two-sided
                                               Incremental           95%            95%
                                           cost-effectiveness     confidence     confidence
Obs   Method                                      ratio            interval       interval

 1    Monte Carlo Interval (Percentile)          3329.07           2214.39        6674.50
 2    Fieller's Theorem                          3329.07           2209.04        6648.17

                           Upper limit    Lower limit
                              of the         of the
       Lower limit of      left-sided     right-sided     Upper limit of
       the left-sided         95%            95%         the right-sided
       95% confidence      confidence     confidence     95% confidence
Obs       interval          interval       interval         interval         Note    NET_COST

 1    Negative Infinity      5727.27        2341.67      Positive Infinity             499.36
 2                               .              .                                      499.36


      SE_NET_     NET_     SE_NET_    Correlation_
Obs     COST     EFFECT     EFFECT         CE         Confidence    Replications

 1     48.91      0.15       0.04         0.33            95           100000
 2     48.91      0.15       0.04         0.33            95           100000
```

Figure S2: Distribution of the simulated ICER from the SAS Macro



## Overview of the Search and Screening Process for Empirical Cost-effectiveness Studies in Education

We conducted a systematic review to identify scholarly publications that reported empirical estimates of the effectiveness and cost of educational interventions. We included studies that i) were randomized control trials (RCTs) in education, ii) collected student academic outcomes for students from grades K to 12, iii) were conducted in a U.S. setting, iv) reported estimates of the effectiveness and cost of the intervention, and v) were published in English. We used a comprehensive search strategy to identify and retrieve all relevant studies. We searched four large databases of indexed research in education and social sciences namely Education Resources Information Center (ERIC), ProQuest Central, ProQuest Central Dissertations and Theses Global, and SCOPUS. In addition, we did forward and backward reference harvesting and Google Scholar random search to locate additional studies. The following search term was used to locate relevant studies: Academic and (achiev* or outcome* or improv*) AND Randomized* and (control* or experiment* or trial* or design or methodology*) AND (cost* or budget* or finance* or expenditure* or expens*) AND (primary or elementary or preschool or "middle school*" or "high school*" or secondary).

The search retrieved 604 total publications in which only 530 publications were unique. Together with 21 publications located through reference harvesting and Google Scholar random search, we had a pool of 551 publications to screen for eligibility. Our screening of the abstract and full text identified 20 publications that meet our criteria for eligibility. The other publications were screened out for various reasons, primarily of which were because the studies were RCTs in health sciences, were not conducted in a U.S. setting or were not targeting K-12 students, did not provide estimates of the effectiveness and/or cost of the intervention, or did not collect student academic outcomes. Of the 20 publications that meet our eligibility criteria, only 10 publications reported the ICERs, in which all the eight publications reported the point estimates (Barret et al., 2020; Barrett & VanDerHeyden, 2020; Clark et al., 2020; Cil et al., 2023; Finster et al., 2023; Guryan et al., 2020; Hollands et al., 2016; Hunter et al., 2018; Kim et al., 2011; Scammacca et al., 2020) and only two of them also reported the confidence interval (Cil et al., 2023; Hunter et al., 2018).

In addition, we conducted a separate search and review for quasi-experimental studies using the same criteria and procedure, except that we focused on quasi-experimental studies instead of RCTs. Therefore, for example, we used the following search term to locate relevant studies: Academic and (achiev* or outcome* or improv*) AND (Quasi-experiment* or propensity score match*) AND (cost* or budget* or finance* or expenditure* or expens*) AND (primary or elementary or preschool or “middle school*” or “high school*” or secondary). The search identified a total of 166 publications including 16 duplicates across the four databases. Our reference harvesting and Google Scholar random search retrieved two additional publications. Of 152 publications screened for eligibility, only five were deemed to meet our review criteria. The vast majority of the studies were excluded mainly because they did not provide cost data, were conducted in a non-US or higher education setting, or were collecting health related outcomes. Of the five publications that meet our eligibility criteria, three publications reported the ICERs, in which all the three publications reported the point estimates (Borman & Hewes, 2002; Bowden & Bellfield, 2015; Bowden et al., 2017) and only one of them also provided the confidence interval (Bowden & Bellfield, 2015).