**DEPARTMENT OF ECONOMICS AND FINANCE**

**SCHOOL OF BUSINESS AND ECONOMICS**

**UNIVERSITY OF CANTERBURY**

**CHRISTCHURCH, NEW ZEALAND**

# Power to the Researchers: Calculating Power After Estimation

**Alex Tian**
**Tom Coupé**
**Sayak Khatua**
**W. Robert Reed**
**Ben Wood**

***WORKING PAPER***

**No. 17/2022**

**Department of Economics and Finance**
**UC Business School**
**University of Canterbury**
**Private Bag 4800, Christchurch**
**New Zealand**

# WORKING PAPER No. 17/2022

## Power to the Researchers: Calculating Power After Estimation

**Alex Tian**[1]
**Tom Coupé**[1]
**Sayak Khatua**[2]
**W. Robert Reed**[1†]
**Ben Wood**[3]

October 2022

**Abstract:** Calculating statistical power before estimation is considered good practice. However, there is no generally accepted method for calculating power after estimation. There are several reasons why one would want to do this. First, there is general interest in knowing whether ex ante power calculations are dependable guides of actual power. Further, knowing the statistical power of an estimated equation can aid one in interpreting the associated estimates. This study proposes a simple method for calculating power after estimation. To assess its performance, we conduct Monte Carlo experiments customized to produce simulated datasets that resemble actual data from studies funded by the International Initiative for Impact Evaluation (3ie). In addition to the final reports, 3ie provided ex ante power calculations from the funding applications, along with data and code to reproduce the estimates in the final reports. After determining that our method performs adequately, we apply it to the 3ie-funded studies. We find an average ex post power of 75.4%, not far from the 80% commonly claimed in the 3ie funding applications. However, we observe significantly more estimates of low power than would be expected given the ex ante claims. We conclude by providing three examples to illustrate how ex post power can aid the interpretation of estimates that are (i) insignificant and low powered, (ii) insignificant and high powered, and (iii) significant and low powered.

**Keywords:** Ex Ante Power, Ex Post Power, Hypothesis Testing, Monte Carlo simulation

**JEL Classifications:** C12, C15, C18

Acknowledgements: We acknowledge helpful comments from David McKenzie, Dorian Owen, Marco Reale, Fabian Dunker, and seminar participants at the BIBaP workshop at the University of New South Wales, the New Zealand Economics eSeminar Series, and Victoria University of Wellington. We also acknowledge 3ie's generous sharing of the data, code, and research applications that enabled our research.

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND
[2] School of Public Policy, Oregon State University, USA
[3] Heifer International, Washington, D.C., USA

† Corresponding author: W. Robert Reed. Email: bob.reed@canterbury.ac.nz

## I. Introduction

Statistical power is the probability that a sample produces a statistically significant estimate given a nonzero effect in the population. Good practice calls for researchers to calculate statistical power when designing experiments. If a study has low power, a researcher may fail to obtain a significant estimate even if a meaningful effect exists. Knowing that a study has low power can be useful even if a study produces a significant estimate. It can alert the researcher that the given effect suffers from publication bias, what Gelman & Carlin (2014) call Type M error.

Calculations of statistical power before estimation is known as ex ante power. There exists a wide variety of methods and procedures for doing this (Huber, 2019; Gertler et al., 2016; Glennerster & Takavarasha, 2013; Coppock, 2013; Djimeu & Houndolo, 2016). However, predicting vital aspects of estimation before the data are seen is speculative business. As Coppock (2013) notes, "in most power analyses you are in fact seeing what happens with numbers that are to some extent made up". As a result, it would be preferable to calculate statistical power after estimation is completed; that is, ex post power. However, to date, there is no generally accepted procedure for doing so.

This study makes two contributions. First, we present a simple method for calculating ex post power that is a variant of previous approaches (Ioannidis et al., 2017; McKenzie & Ozier, 2019). We demonstrate its usefulness through a series of Monte Carlo experiments to assess its performance. The experiments are designed to be similar to "real-world" data from projects funded by the International Initiative for Impact Evaluation (3ie).

Our second contribution consists of two applications of our method. Through an agreement with 3ie, we were able to access research materials for 23 studies funded by 3ie. 3ie supplied the original funding applications, including ex ante power calculations,

along with the final report, data, and statistical code. This allowed us to compare our estimates of ex post power with researchers' ex ante calculations for 47 estimated treatment effects drawn from the 23 studies. We observe an average ex post power of 75.4%, not far from the 80% commonly claimed in the 3ie funding applications. However, a disproportionate number are low-powered estimates; more than would be expected if all studies had 80% true power. Using regression analysis, we find that most of the differences between ex post and ex ante power (58%) can be explained by differences in planned and actual total observations, number of clusters, and intracluster correlation (ICC).

Our other application uses three examples to illustrate how our method can be used to assess individual estimates. The first two examples feature statistically insignificant estimates from 3ie's sample of studies. We show how calculation of ex post power can provide insight into whether an insignificant estimate is due to a negligible treatment effect, or, alternatively, a research design that is insufficiently powered. Our third example consists of a statistically significant estimate. We show how calculation of ex post power can alert researchers to the possible seriousness of Type M error (Gellman & Carlin, 2014) whereby significant estimates overstate the size of the population effect.

Our study proceeds as follows. Section 2 provides a brief introduction to the subject of statistical power. Section 3 reviews the literature on ex post power and presents our method, something we call the SE-ES method. Section 4 discusses the design of the Monte Carlo experiments used to assess the performance of the SE-ES method. Section 5 reports the associated results. Sections 6 and 7 provide two applications of the SE-ES method to 3ie-funded impact evaluations. Section 8 concludes.

## 2. Calculation of Power

Statistical power is a function of relatively few parameters. Let $ES$ be the population effect size of a given treatment; $\widehat{ES}$ a sample estimate of $ES$; $s.e.(\widehat{ES})$ the population standard deviation of the distribution of sample estimates, $\widehat{ES}$; and $\widehat{s.e.(\widehat{ES})}$ a sample estimate of $s.e.(\widehat{ES})$. Define

(1.a) $t_{ES} \equiv \frac{ES}{s.e.(\widehat{ES})}$,

and

(1.b) $\hat{t}_{ES} \equiv \frac{\widehat{ES}}{\widehat{s.e.(\widehat{ES})}}$.

If $\hat{t}_{ES}$ is distributed according to Student's $t$ distribution with $\upsilon$ degrees of freedom, then power is calculated by

(2) $t_{\upsilon,1-Power} = \left(t_{\upsilon,1-\alpha/2} - t_{ES}\right)$.

In other words, given values for (i) the effect size, $ES$; (ii) significance level, $\alpha$; and (iii) degrees of freedom, $\upsilon$; one can calculate $Power$ as a function of the population parameter, $s.e.(\widehat{ES})$.

FIGURE 1 illustrates the relationship between $Power$ and $s.e.(\widehat{ES})$. We set $\alpha = 5\%$ and $\upsilon = 50$. Accordingly, $t_{50,0.975} \approx 2$. For positive values of $ES$, $Power$ is the probability that $\hat{t}_{ES} > 2$.[1] FIGURE 1 shows three cases: (i) $t_{\upsilon,1-\alpha/2} > t_{ES}$, (ii) $t_{\upsilon,1-\alpha/2} = t_{ES}$, and (iii) $t_{\upsilon,1-\alpha/2} < t_{ES}$. We fix $ES = 4$ in all three cases and reduce the value of $s.e.(\widehat{ES})$ in steps from 3 to 2 to 1.5. With each step, the distribution of $\hat{t}_{ES}$ shifts to the right.

[1] Strictly speaking, it is the sum of the probabilities that $\hat{t}_{ES} > 2$ and $\hat{t}_{ES} < -2$. However, as a practical matter, except for very small values of $|t_{ES}|$, only one of the tails of the distribution of $\hat{t}_{ES}$ has a non-negligible probability.

For example, when $t_{\upsilon,1-\alpha/2} = t_{50,0.975} = 2$, and $t_{ES} = \frac{4}{3} = 1.33$, then $\left(t_{\upsilon,1-\alpha/2} - t_{ES}\right) = 0.67$. The value of $Power$ that makes $t_{\upsilon,1-Power} = 0.67$ is 0.253, because $t_{50,0.747} = t_{50,1-0.253} = 0.67$. As $s.e.(\widehat{ES})$ decreases to 2 and then 1.5, $Power$ increases to 0.500 and 0.747. In words, as the estimates of $ES$ becomes more precise, a larger percentage of estimates will be statistically significant.

An important insight from Equation (2) is that it highlights that analytic methods for calculating ex ante power essentially consist of forecasting $s.e.(\widehat{ES})$. Programs like Stata's *power* command and the free software program *G*Power* (Faul et al., 2007) require the user to input various aspects of the data, such as the sample size, standard deviation of the output variable, percent of variation in the dependent variable explained by covariates, percent of observations receiving treatment, number of clusters, and the intra-class correlation (if relevant). These inputs are combined to produce an estimate of $s.e.(\widehat{ES})$, which in turn is used to calculate power. It bears emphasizing that the requisite inputs are supplied before one actually sees the data.

## III. Literature review of ex post power and description of the SE-ES method

"Observed power". One method that has been commonly used in the past for calculating power after estimation ("ex post power") is often referred to as "observed power". It uses both the estimated effect size, $\widehat{ES}$), and its associated estimated standard error, $\widehat{s.e.(ES)}$. Specifically, it replaces $t_{ES}$ with $\hat{t}_{ES}$ in Equation (2). This approach is now widely recognized as flawed (Hoenig & Heisey, 2001; Yuan & Maxwell, 2005). It produces a biased estimate of true $Power$ when true $Power \neq 50\%$. Arguably worse, it produces highly imprecise estimates. This is illustrated by three examples in FIGURE 2.

When true $Power = 50\%$, "observed power" is uniformly distributed between 0% and 100%. This is illustrated in the top panel. While "observed power" is unbiased in this

case, it is very imprecise. When true $Power < 50\%$, not only is "observed power" an imprecise estimator, it is also upwardly biased. The middle panel of FIGURE 2 shows the distribution of "observed power" when true $Power = 20\%$. It has a mean of 27.5%. When true $Power > 50\%$, "observed power" is biased in the opposite direction. The bottom panel shows the associated distribution when true $Power = 80\%$. It has a mean of 72.3%. As these deficiencies have become recognized, "observed power" has fallen into disfavor.

Ex ante methods applied to ex post data. Another approach to estimating ex post power uses ex ante methods where the inputs consist of characteristics from the final dataset. An example of this is Skiba & Tobacman (2019). To demonstrate that their empirical analysis is sufficiently sized, they calculate sample sizes necessary to achieve 80% power when estimating economically important effect sizes. They do this using Stata's *power* command, which is commonly used for ex ante power analyses. While they solve for sample sizes, they could just as easily have supplied their actual sample sizes to calculate ex post power. The point is, they use the same formulae/software that are used for ex ante power calculations to address power concerns after estimation. Or to state it differently, they input sample characteristics to predict $s.e.(\widehat{ES})$ rather than using direct estimates of $s.e.(\widehat{ES})$ from the final estimating equation.

Bootstrapping procedures. Kleinman & Huang (2017) propose a bootstrapping procedure for estimating power for a binary treatment. Bootstrapping is related, but different, from simulation. Simulation uses sample characteristics to build artificial datasets that resemble the data on which final estimation will be done. In contrast, bootstrapping builds artificial datasets using the data itself. While K&H's method can only be applied to pre-treatment data, it illustrates the bootstrapping approach.

Pre-treatment data are randomly assigned to treatment and control groups. The control group is resampled with replacement to create artificial control datasets. The treatment group is also resampled with replacement, but a hypothesized treatment effect is added to each observation's output variable. Treatment and control datasets are then matched and tested for differences. The bootstrapped power equals the percent of tests in which the null hypothesis is rejected.

Brown, Lambert, & Wojan (2019) recently proposed another bootstrapping procedure that is specifically designed to be applied post-estimation. Further, it easily handles continuous treatments. They take coefficient estimates and residuals from the final estimating equation. They then construct a parent, artificial dataset, replacing the estimated treatment effect with a hypothetical treatment effect. New values of the output variable are created by combining predictions from the edited regression specification with the original residuals. Paired bootstrapping is applied to this parent dataset to produce multiple artificial datasets. Estimation is then carried out on the individual, artificial datasets. Ex post power is calculated as the percent of times the null hypothesis is rejected.

Methods that rely on the estimated standard error. As noted above, other than "observed power", none of the methods above use the estimate of $s.e.(\widehat{ES})$ that comes from the final estimating equation. Two recent approaches do this. Ioannidis, Stanley, & Doucouliagos (2017) combine meta-analysis with the estimated standard error to produce ex post power estimates. They reach the following startling conclusion: "We survey 159 empirical economics literatures that draw upon 64,076 estimates of economic parameters reported in more than 6,700 empirical studies. Half of the research areas have nearly 90% of their results under-powered. The median statistical power is 18%, or less."

Essentially, IS&D substitute $\frac{\widehat{ES}_{Meta-analysis}}{\widehat{s.e.(ES)}}$ for $t_{ES}$ in Equation (2). That is, they substitute the estimate of $s.e.(\widehat{ES})$ from individual studies for the population parameter, $s.e.(\widehat{ES})$; and the literature-wide, overall average estimated effect, $\widehat{ES}_{Meta-analysis}$ for $ES$. This allows them to calculate ex post power for each of the 64,075 estimates in their sample. While useful for calculating literature-wide estimates of overall power, IS&D's approach isn't applicable to single studies. It requires an estimate from a meta-analysis and thus cannot be applied to a stand-alone, individual estimate.

In contrast, McKenzie & Ozier (2019) propose an alternative method that is applicable to single studies. The main difference to what we propose in this paper is that they use Equation (2) to calculate the value of $ES$ that can be estimated with a given value of $Power$:[2]

(3) $ES = \left(t_{v,1-\alpha/2} - t_{v,1-Power}\right) \times \widehat{s.e.(ES)}$.

Note that everything on the right-hand-side of the equation is either a given parameter $(v, \alpha, Power)$ or comes from the equation after it has been estimated $(\widehat{s.e.(ES)})$.

The Standard Error-Effect Size (SE-ES) method. A rearrangement of terms in Equation (3) produces the following equation for estimating ex post power for (i) a given effect size, $ES$; (ii) estimated standard error, $\widehat{s.e.(ES)}$; and (iii) parameter values $v$ and $\alpha$:

(4) $t_{v,1-Power} = t_{v,1-\alpha/2} - \frac{ES}{\widehat{s.e.(ES)}}$.

We call this the Standard Error-Effect Size (SE-ES) method. Unlike "observed power", the SE-ES method uses a hypothesized value for ES rather than an estimated value. It is

[2] This is commonly known as the Minimum Detectable Effect (MDE).

applicable anytime an estimation procedure produces an estimate of the standard error, and the ratio $\frac{\widehat{ES}}{s.e.(\widehat{ES})}$ is distributed according to Student's $t$ distribution with $v$ degrees of freedom. When finite sample statistics are not applicable, $t_{v,1-Power}$ and $t_{v,1-\alpha/2}$ can be replaced by $z_{1-Power}$ and $z_{1-\alpha/2}$, respectively, where $z$ is distributed standard normal.

In addition to being simple to apply, Equation (4) has the added benefit of easily accommodating alternative estimation procedures. The formula is accurate for OLS, cluster-robust OLS, IV estimation, and alternative maximum likelihood procedures, among others. All that is required is that the estimate of the standard error be an accurate measure of the true variability of $\widehat{ES}$. This will not always be the case. For example, it is well known that conventional estimates of cluster-robust standard errors may underestimate true variability when the number of clusters is small (MacKinnon, 2019; Roodman et al., 2019). In these cases, using conventional sandwich estimators to estimate standard errors will result in biased ex post power estimates.

As we saw in the case of "observed power" when true $Power = 50\%$, just being unbiased is insufficient to make an estimator useful. We also want it to be precise. On this count, both Ioannidis, Stanley & Doucouliagos (2017) and McKenzie & Ozier (2019) provide little help.[3] While they apply their methods to estimate ex post $Power$ and ex post MDE, respectively, there is little evidence to assess their reliability. Accordingly, our

[3] McKenzie and Ozier (2019) do provide simulation evidence, but it focuses on MDE, not *Power*. Further, since it is a blog and not an academic article, it only provides minimal evidence of the reliability of their method.

next order of business is to assess the performance of the SE-ES method as an estimator of $Power$.[4]

## IV. Design of the Monte Carlo Experiments

In this section we discuss the design of the Monte Carlo experiments we use to assess the performance of the SE-ES power estimator. Since the goal is to apply this estimator to our sample of 3ie-funded projects, we first describe those projects and identify key data characteristics that we want to incorporate in the design of the experiments.

Introduction to 3ie. 3ie is a non-governmental funding agency whose primary mission is to support impact evaluations and systematic reviews of programs to help the poor in low- and middle-income countries. They were founded in 2008 and have an annual budget of approximately $25 million (USD).

From its beginning, 3ie recognized the importance of transparency and research quality. As a result, they implemented policies that enabled a greater level of data quality assurance than typical review processes for journals or institutional donors. Over time, they required highly detailed pre-analysis plans, evaluation registration, and survey data sharing. Upon completion, as a contingency of their funding, researchers provided 3ie with the data supporting their analyses. After removing any information that allowed personal identification, 3ie warehoused these data for later, secondary research.

As part of the application process, 3ie required applicants to provide ex ante power calculations to ensure that their evaluation had sufficient power to identify economically important treatment effects. This was one of the criteria 3ie used to assess the strength of the research applications they received. After the evaluation concluded,

[4] Tian (2021) compares the SE-ES method to the bootstrapping procedure of Brown, Lambert, & Wojan (2019). He finds that the SE-ES performs slightly better on the dimension of mean squared error. However, its major advantage lies in its ease of application as it does not require the construction of thousands of artificial datasets.

the researchers submitted their reports to 3ie along with the data behind their analysis. Over time, these requirements expanded to require researchers to also include statistical code that allowed for push-button replication (3ie(a), n.d.) These policies eventually resulted in 3ie's Transparency, Research, and Ethics policy (3ie(b), n.d.).

Through an agreement with 3ie, we were given access to the full files for 23 impact evaluations. As a result, we not only have the final estimates, with the respective standard errors, but we also know the hypothetical effect sizes that were used to calculate ex ante power. Further, we have the data itself that allows us, in most cases, to identify differences between planned and actual values of sample size, number of clusters, ICC values, and other data characteristics important for statistical power.

Our confidentiality agreement with 3ie prevents us from revealing the identity of the impact evaluations included in our sample. However, the following list, taken from 3ie's website (3ie(c), n.d.), gives an idea of the kinds of studies that 3ie funds. Note that none of these studies are included in our sample.

- "Community advocacy forums and public service delivery in Uganda: Impact and the role of information, deliberation and administrative placement"
- "Evaluation of secondary school teacher training under the School Sector Development Programme in Nepal"
- "Impacts of supportive feedback and nonmonetary incentives on child immunisation in Ethiopia"
- "Impacts of electronic case management systems on court congestion in the Philippines"
- "Impacts of the Stimulate, Appreciate, Learn and Transfer community engagement approach to increase immunization coverage in Assam, India"
- "Impacts of a novel mHealth platform to track maternal and child health in Udaipur, India"
- "Impacts of engaging communities through traditional and religious leaders on vaccination coverage in Cross River State, Nigeria"

Many of the studies in our sample contained more than one estimated treatment effect. We chose as many as could be matched to initial ex ante power calculations. Our final dataset consists of 47 estimated treatment effects from the 23 studies.

What do the 3ie datasets look like? As much as possible, we want our artificial datasets to look like actual 3ie datasets. Three important data characteristics are sample size, number of clusters, and ICC values. FIGURE 3 reports histograms for all three characteristics for the 47 datasets associated with the estimated treatment effects from the 3ie projects.

Based on the first two histograms in FIGURE 3, we create simulated datasets where total observations take one of two values – 3,000 and 10,500; and clusters take one of four values: 60, 100, 150, and 250. The respective values are indicated in the histograms by dashed, vertical red lines. Let $N$ represent the number of clusters, $T$ the number of subjects within a cluster, and let all the simulated datasets be balanced. Then the values above define four basic dataset configurations, $N \times T =$ (i) $60 \times 50$, (ii)$100 \times 30$, (iii)$150 \times 70$, and (iv) $250 \times 42$, where the respective datasets have either 3,000 or 10,500 total observations.

Design of the Monte Carlo Experiments. Our experiments specify the following linear data generating process (DGP):

$$y_{nt} = 1 + ES_{Power} \times T_{nt} + \varepsilon_{nt}, \quad (5)$$

where $y_{nt}$ is the output of interest, $T_{nt}$ is a binary treatment variable, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega})$, and the subscripts $n$ and $t$ represent the individual cluster and subject indicators, $n = 1,2, \dots N$, $t = 1,2, \dots, T$. Half of the clusters receive treatment, and everybody in a treated cluster receives the treatment. This fully specifies the $(NT \times 2)$ data matrix $\boldsymbol{X} = [\boldsymbol{i} \quad \boldsymbol{T}]$, where $\boldsymbol{i}$ is a column vector of ones and $\boldsymbol{T}$ is a column of half ones and half zeroes . $ES_{Power}$ is the population treatment effect. We set the value of $ES_{Power}$ such that the associated effect

size corresponds to a specific power value, $Power = (10\%, 20\%, 30\%, \ldots, 80\%, 90\%)$. To do that, we need to specify $\boldsymbol{\Omega}$.

The bottom panel of FIGURE 3 reports a histogram of ICC values for the 3ie datasets. Based on this, we select three ICC values to be representative of the full sample: $\rho = 0.050$, 0.150, and 0.250. The respective error variance-covariance matrices (VCMs) are composed of $N^2$ $(T \times T)$ blocks as configured below, with all the blocks on the main diagonal consisting of 1's and $\rho's$, and 0's everywhere else:

$$(6) \qquad \boldsymbol{\Omega}_{NT\times NT} = \sigma^2 \times \begin{bmatrix} 1 & \rho & \cdots & \rho & & 0 & 0 & \cdots & 0 \\ \rho & 1 & \cdots & \rho & & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 & & 0 & 0 & \cdots & 0 \\ & & \vdots & & \ddots & & & \vdots & \\ 0 & 0 & \cdots & 0 & & 1 & \rho & \cdots & \rho \\ 0 & 0 & \cdots & 0 & & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & & \rho & \rho & \cdots & 1 \end{bmatrix},$$

The three ICC values ($\rho$) in combination with the four basic dataset configurations ($N \times T$) generate twelve unique error VCMs. Without loss of generality, we let $\sigma^2 = 1$.[5]

To be consistent with the 3ie studies in our sample, the Monte Carlo experiments estimate the treatment effect using OLS with cluster-robust standard errors. The associated VCM for the estimated coefficients in Equation (5) is given by:

$$(7) \qquad Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

where the second element in the $2 \times 1$ vector $\widehat{\boldsymbol{\beta}}$ is $\widehat{ES}$. Note that everything on the right-hand-side is known. Thus, $s.e.(\widehat{ES})$, which is the square root of the lower right element of $Var(\widehat{\boldsymbol{\beta}})$, is easily solved as a function of population parameters and known constants.

Given $s.e.(\widehat{ES})$, it is straightforward to solve $ES_{Power}$ for any given $Power$ value:

$$(8) \qquad ES_{Power} = \left(t_{v,1-\alpha/2} - t_{1-Power}\right) \times s.e.(\widehat{ES}).$$

[5] $\sigma^2$ is a nuisance parameter because it merely scales the size of $ES_{Power}$.

As a result, we have everything we need to generate artificial datasets following the DGP in Equation (5). The 9 different $Power$ values, $Power = (10\%, 20\%, 30\%, \ldots, 80\%, 90\%)$, in combination with 12 dataset configurations/error variance covariances $(N \times T, \boldsymbol{\Omega})$, produce a total of 108 experiments. Each experiment consisted of 1000 replications.

## V. Results of Monte Carlo Experiments Assessing the Performance of the SE-ES Estimator

Explanation of the tables. The results of the Monte Carlo experiments are reported in TABLES 1 and 2. TABLE 1 reports the benchmark results for OLS assuming no intra-cluster correlation $(\rho = 0)$. The estimated OLS standard errors are robust for heterokedasticity. TABLE 2 does the same for the clustered data, $\rho = 0.050, 0.125, 0.250$, where the estimated standard errors are robust for both heteroskedasticity and intra-cluster correlation.

TABLE 1. For each experiment, we report the following summary statistics for the sample of 1000 ex post power estimates: (i) the mean, (ii) the lower ("p(0.05)") and (iii) upper bounds ("p(0.95)") of a 90% sample interval, and (iv) the standard deviation. The table is organized in nine panels, corresponding to the nine $Power$ values. The rows of each panel report estimates for sample sizes of 3,000 and 10,500, respectively.

The values in the table are easily misinterpreted. For example, when $Power = 10\%$ and *Sample Size* = 3,000, the 90-percent sample interval of estimated power values ranges from 9.7% to 10.3%. When *Sample Size* increases to 10,500, the associated interval narrows to 9.9% and 10.1%. Note that *Power* is held constant at 10% in both cases. To maintain *Power* as the sample size increases, the associated effect size, $ES_{Power}$, is reduced. Thus, the narrower range of estimated power values is not due to the fact that $s.e.(\widehat{ES})$ is smaller for the larger sample. This has already been accommodated by the

smaller $ES_{Power}$ value. Instead, it is due to the fact that $s.\widehat{e.(\widehat{ES})}$ is a more precise estimator of $s.e.(\widehat{ES})$.

Overall, the values in the table indicate a high level of reliability. The SE-ES ex post power estimates are unbiased at all levels of *Power*. The 90-percent sample intervals are all relatively narrow. When sample size = 3,000, the intervals all lie within 2 percentage points of their true *Power* values. For example, for $Power = 50\%$, the 90-percent sample interval is (48.4%, 52.0%). When sample size = 10,500, the intervals all lie within 1 percentage point of their true *Power* values. For true $Power = 50\%$, the corresponding sample interval is (49.1%, 50.9%).

TABLE 2. TABLE 2 reports results for the more representative case with clustering. These results are directly applicable to the 3ie-funded studies, as all 23 studies/47 treatment effects were estimated using OLS with cluster robust standard errors. As before, the table is organized in nine panels for the different true *Power* values. The first six rows of each panel report estimates for sample sizes of 3000, and the next six rows report estimates for samples sizes of 10,500. Within each set of six rows, the first three rows report results for the smaller number of clusters (60 versus 100, and 150 versus 250, respectively). Thus, within each set of six rows, one can see the effect of increasing the number of clusters while holding constant the total number of observations. And within each set of 3 rows, one can see the effect of increasing ICC holding constant sample size and the number of clusters.

As noted above, one must be careful to avoid misinterpretation. For example, when $Power = 10\%$, $N = 60$, and $T = 50$, there is hardly any change in the performance of the SE-ES estimator for different values of $\rho$ (≡ ICC). This does not mean that ex post power is unaffected by changes in $\rho$: As $\rho$ increases, $ES_{Power}$ is adjusted to compensate for the fact that $s.e.(\widehat{ES})$ is larger. Rather, differences in sample intervals are related to

the precision of the standard error estimates. Thus, the narrower ranges of the 90-percent sample intervals one observes as the number of clusters increases is due to the fact that $s.\widehat{e.(ES)}$ is a more precise estimator of $s.e.(\widehat{ES})$ when there are more clusters.

We make two general observations from TABLE 2. First, the SE-ES estimator is slightly biased upwards for smaller numbers of cluster. The bias is largest when true *Power = 50%*. When the number of clusters = 60, average estimated power is approximately 51.8%. This falls to approximately 50.5% when the number of clusters increases to 250. This is consistent with the fact that the conventional cluster robust estimator underestimates standard errors when the number of clusters is relatively small (Cameron, Gelbach, & Miller, 2008).

The other observation relates to the sample intervals of the estimated power values. Here, performance is also worst for relatively small numbers of clusters. For example, when the number of clusters = 60 and *Power* = 20%, the lower and upper bounds of the 90-percent interval of estimated power values are approximately 16% and 27%. When *Power* = 50%, the corresponding values are 40% and 65%. And when *Power* = 80%, they are 68% and 92%. In contrast, when the number of clusters = 250, the corresponding intervals are: (*Power* = 20%) 18% to 23%; (*Power* = 50%) 45% to 57%; and (*Power* = 80%), 74% to 86%.

Performance assessment. TABLES 1 and 2 make clear that the performance of the SE-ES estimator depends on the dataset and the estimator used for the standard error. When independence of observations describes the researchers' dataset, so that conventional OLS with heteroskedasticity-robust standard errors is appropriate, the SE-ES estimator performs very well. Across all *Power* values, estimates of ex post power are unbiased and very close to their true values.

When observations are clustered, performance of the SE-ES estimator declines, especially for relatively small numbers of clusters. This is a function of the well-known shortcomings of the conventional, cluster-robust estimator (Cameron, Gelbach, & Miller; 2008; MacKinnon, 2019; Roodman, 2019). We focus on the clustered case because this is the estimator exclusively used by the studies in our 3ie sample. The question is whether the SE-ES estimator is sufficiently precise to be a useful estimator of ex post power for this sample.

In assessing the suitability of the SE-ES estimator, we note that most of the 3ie-funded studies claimed to have 80% power in their funding applications. Therefore, we are most interested in the TABLE 2 results for true $Power = 80\%$. With respect to bias, there is next to no bias in the associated experimental estimates. With respect to precision, the results are harder to assess. If we take the $(N \times T = 60 \times 50 = 3000)$ experiments as a worst or close to worst case scenario for the SE-ES estimator, the lower and upper bounds of the 90-percent sample interval are 68% and 92%. If we jump down to the bottom three rows of the panel, the corresponding values are 74% and 86%. Compared to the "observed power" simulations of FIGURE 2, these results demonstrate immense improvement.

Whether they are good enough to be a suitable estimator of ex post power is a subjective decision. Our judgment is that they are. Approximately 90% of the SE-ES power estimates lie within $\pm 10$ percentage points of 80%, the true power. As we illustrate below, for many purposes, this range is sufficient. What matters is not whether power is 70% or 80% or 90%, but that it is not 30% or 40% or 50%. In judging the value-added of ex post power estimates, we are mindful that ex ante power values are also estimates, based on educated guesses about sample characteristics from as-yet-unseen datasets. Further, the formulae that transform those educated guesses to ex ante power

values are only approximations of $s.e.(\widehat{ES})$. For all these reasons, we believe that the SE-ES method provides a superior approach to measuring the true power of the 3ie estimates.[6]

## VI. Application One: Ex Post Versus Ex Ante Power

Comparison of ex post versus ex ante power (averages). Our first application compares the SE-ES estimates of ex post power with the ex ante values reported in the funding applications. TABLE 3 reports side-by-side ex ante and ex post power values, along with planned and actual sample characteristics for number of clusters, ICC, and total observations, for each of the 47 estimated treatment effects. The last row reports the sample averages.

The average pre- and post-estimation values correspond quite closely. The average number of planned clusters is 151.8, compared to an average of actual clusters of 145.2. The pre-estimation average estimate of ICC is 0.123. Actual average ICC is 0.128. The average, planned sample size for the 3ie studies is 5,883, compared to an average, actual sample size of 5,991. And finally, the average, ex ante power across all 47 treatment effects is 80.8%, compared to an average, ex post power of 75.4%. Thus, based on averages, the SE-ES estimates of power after estimation correspond closely to the ex ante estimates submitted to 3ie.

Comparison of ex post versus ex ante power (individual estimates). The average values above mask a significant amount of heterogeneity at the individual level. FIGURE 4 provides a closer look. The top panel of FIGURE 4 presents a histogram of the individual

[6] The simulations in TABLES 1 and 2 assume that the estimator correctly specifies the error VCM. But suppose it doesn't? The APPENDIX reports simulation results when OLS with robust cluster standard errors is used with clustered data where the errors are cross-sectionally correlated. This violates the assumption of independence of clusters underlying the robust estimator of the error VCM. However, as the Appendix shows, this violation has little effect on the performance of the SE-ES power estimator. In fact, in a wide variety of Monte Carlo experiments that we conducted, we find that the impact of violating the assumption of independence of clusters is negligible.

ex post power estimates. While the average is close to 80%, the individual values range from a minimum of 20.7% to a maximum of 100%. The bottom panel of FIGURE 4 gives a somewhat different look. It compares the ex ante and ex post power estimates, ranked from lowest ex post power estimate to largest. Note that a few of the 3ie studies had ex ante estimates of power greater than 80% (cf. IDs 19, 20, and 22).

If we are willing to use the Monte Carlo experiments as a guide, we can determine whether the observed distribution of ex post power values in FIGURE 4 is consistent with all studies having a true power of 80%. TABLE 2, Panel "True Power = 80%" reports 90% power intervals for the twelve data environments selected to be representative of the 3ie-funded studies. These range from (68.0%, 91.5%) for $(N, T, \rho) = (60, 50, 0.050)$ to (74.6%, 85.6%) for $(N, T, \rho) = (250, 42, 0.250)$.

If true $Power = 80\%$, we would generously expect 5% of the ex post power estimates in TABLE 3 to be less than 65%, or approximately 3 out of the 47. In fact, there are 12 ex post power estimates less than 65% (25.5%), for an "excess" of 9 (out of 47). While this is only a back-of-the-envelope calculation, it does suggest that there is a significant minority of studies that fail to achieve planned power in their final estimating equations.

Determinants of individual differences between ex post and ex ante power. Given that ex ante power calculations are a function of effect sizes and sample characteristics, and given that our power calculations use the same effect sizes, any differences between ex post and ex ante power should be due to difference between planned and actual sample characteristics. The supplementary materials from 3ie provide information on planned total observations, planned number of clusters, and assumed ICC values.

FIGURE 5 compares these with actual values taken from the final estimating equations. We report differences in units of percent for total observations and clusters,

and straight differences for ICC. In most cases, planned and actual values are reasonably close. However, in a few cases the differences are quite large. For example, one of the studies had planned to have 4000 observations, but the final estimating equation only used 831 observations. Another study had planned to have 60 clusters, but the final dataset only had 20. A third study very conservatively planned for an ICC of 0.500, but the final dataset was characterized by a much lower level of intracluster correlation (0.060).

TABLE 4 investigates the extent to which these sample characteristics can explain the differences between ex post and ex ante power. We estimate four specifications where the dependent variable is the difference between ex post and ex ante power, and the explanatory variables consist of various combinations of differences in total observations (in percent), clusters (in percent), and ICC (straight difference). Specification (1) is the baseline specification consisting of the three difference variables. Specification (2) adds quadratic terms for all three variables to the baseline specification. Specification (3) adds interaction terms to the baseline specification. And Specification (4) adds both quadratic and interactions terms to the baseline specification.

From the baseline specification in (1), we estimate that a 10 percent increase in the difference between actual and planned sample size (ΔObs), holding constant the difference in clusters and ICC, increases the difference between ex post and ex ante power (ΔPower) by 1.4 percentage points. In contrast, a 10 percent increase in the number of actual clusters over planned clusters (ΔClust), holding constant the difference in total observations and ICC, is estimated to increase ΔPower by 2.6 percentage points. And a 0.10 increase in ICC, holding the other difference variables constant, reduces ΔPower by 3.5 percentage points. Of these, only the estimate for the difference in clusters is

significant at the 5 percent level. This basic specification "explains" approximately 39% of the variance in ΔPower.

The other specifications add different combinations of quadratic and interaction terms. The "best" specification according to the BIC is Specification (2), which consists of linear and quadratic terms of the three difference variables, but no interaction terms. The inclusion of the quadratic terms is further supported by the fact that they are jointly significant at the 5 percent level. This expanded specification "explains" over half of the observed differences between ex post and ex ante power (58%).

TABLE 4 highlights the importance of being able to correctly predict the sample characteristics of the final dataset when planning the research design. The histograms from FIGURE 5 illustrate just how difficult this is. Taken together, these results underscore the importance of being judiciously sceptical of ex ante calculations of power. They also underscore the potential benefit of estimating ex post power.

## VII. Application Two: Using ex post power to interpret estimates from individual studies

Perhaps the most practical application of the SE-ES method is in providing guidance when interpreting individual estimated effects after estimation. This section provides three examples to illustrate this. The first two examples illustrate how ex post power estimates can inform interpretation of statistically insignificant estimates. The last example shows how ex post power can be used to assess significant estimates.

Insignificant with low power. The first example comes from ID =13. This project assessed the impact of a government family planning program. Ex ante power calculation focused on a binary outcome variable indicating utilization of services. Treatment consisted of a variety of outreach activities in selected communities to encourage people to use family planning services. The treatment variable was also a binary variable that indicated that the subject resided in a community receiving treatment. The ex ante power

calculations assumed an effect size of 0.060 to correspond with 80% statistical power ($ES_{Power} = 0.060$). The actual estimate reported in the final study was 0.077 with an estimated (cluster robust) standard error of 0.051. Despite being larger than the assumed effect size, the estimated treatment effect was not statistically different from zero at the 5% level.

A common interpretation (misinterpretation) of statistical insignificance would see this result as evidence that the outreach activities were ineffective in achieving their objective of getting people to access family planning services. The problem with this interpretation is that it ignores the power of the respective estimating equation. The SE-ES method estimates that this equation only had 20.7% power for an effect size of 0.06. In other words, the researchers only had a one in five probability of obtaining a significant estimate given their assumed treatment effect.

This example shows how ex post power can address the question, Is an insignificant estimate due to a negligible effect or insufficient power? Assuming 0.06 was the true effect, there was little likelihood of the researchers obtaining a statistically significant estimate. In this particular case, the problem was not sample design. The problem was that the actual number of clusters (63) fell far short of what was planned (118). The small number of clusters reduced precision, and this in turn reduced power. Knowledge of ex post power helps one guard against misinterpretation of statistically insignificant estimates.

Insignificant with high power. A second example comes from ID =1. This program funded youths to start their own businesses. The outcome variable measured respondents' assets. The treatment variable was binary and consisted of assignment to a group eligible to receive funding. Estimation was based on a difference-in-differences (DID) specification using OLS.

The authors' based their calculation of 80% ex ante power assuming a Cohen's $d$ value of 0.2,

$$(9) \qquad d = \frac{|\bar{y}_T - \bar{y}_C|}{S_{pooled}},$$

where $\bar{y}_T$ and $y_C$ are means of the outcome variable for the treatment and control groups, and $S_{pooled}$ is the pooled sample standard deviation of the outcome variable. The latter is equivalent to the root mean squared error (RMSE) in a simple regression with a dummy variable for treatment. Cohen's $d = 0.2$ is widely taken as representing a "small effect" (Cohen, 1992).[7] Using the data provided by 3ie, we estimated a RMSE $= S_{pooled} = 5.889$, which implies an estimated treatment effect in the OLS-DID specification of $ES_{Power} = 1.1778$. In the final report, the authors reported an estimated treatment effect of 0.455 with a standard error of 0.385 and a p-value of 0.238. Accordingly, we estimate the power of the final estimated equation to be 86.3%.

This is the opposite case of the previous example. Here we have a statistically insignificant estimate with relatively high power. If the true effect had been equal to or greater than 1.1778 (equating to a "small effect" by Cohen's metric), the probability of obtaining a significant estimate would have been equal to or greater than 86.3%. The fact that the estimated treatment effect proved to be insignificant suggests that the true effect is less than 1.1778. Thus, knowing that the final estimating equation had high power allows one to identify negligible effect as the likely cause of the statistical insignificance.

Significant with low power. Our last case illustrates how ex post power can be used to assist interpretation of statistically significant estimates. Unfortunately, our sample of 3ie-funded studies did not produce an example where the estimated effect was

[7] While it quite common to assume that Cohen's $d = 0.2$ is "small", this depends on the particular problem being examined. An intervention that increased the dependent variable by 0.2 standard deviations could, in some circumstances, be large; even unreasonably large. We thank David McKenzie for pointing this out.

statistically significant with low power. However, study ID = 18 is close enough to illustrate the point. This study evaluated a program designed to reduce conflict between majority and minority ethnic group youths. The intervention consisted of a 6-8 week, extra-curricular course administered in public schools. Students were randomly assigned to treatment and control groups. The outcome variable was a Likert scale response variable that measured trust in new people they meet on a scale from 1 (do not trust at all) to 4 (trust completely).

Similar to the previous example, the authors based their power calculation of 80% assuming a Cohen's *d* value of 0.24. This converts to an effect size in units of the original variable of 0.191. The authors used a DID specification in an OLS regression equation and estimated a treatment effect of 0.273 with a standard error of 0.101 and a "marginally significant" p-value of 0.051. Based on the SE-ES method, we estimate that this equation had power of 37.0%.

In other words, if the population effect size had been 0.191, there is less than a 40% chance that this study would have produced a significant estimate. In fact, the 3ie-funded study reported an estimate of 0.273 with a p-value of 0.051. One interpretation of these results is that the effect was equal to 0.191 and the authors were relatively lucky to obtain a marginally significant estimate. A second possibility is that the true effect size was larger than 0.191 which is why the study estimated the larger estimate.

However, a third possibility is that this is a case of Gelman & Carlin's (2014) Type M error. Obtaining a marginally significant estimate when one would otherwise not expect one is an indicator that the estimate may an outlier in terms of both statistical

significance and magnitude. Low, ex post power can be a reason to downgrade one's confidence in an estimated result.[8]

## VIII. Conclusion

While it is generally recognized that researchers should do power calculations before estimation ("ex ante power"), to date there is no generally accepted method for calculating power after estimation ("ex post power"). A method to calculate ex post power would be useful for multiple reasons. For one, it would be enlightening to know whether ex ante power calculations were generally reliable. For another, knowing the power of an estimating equation could help one interpret the results. For example, it could help one to know whether an insignificant estimate was due to the underlying effect being small, or because the study was underpowered.

This study introduces a simple method for calculating ex post power that we call the SE-ES method. We then conduct a series of Monte Carlo experiments to assess its performance. We find that performance depends on both the estimator being used (for example, whether one uses heteroskedasticity-robust estimates of the standard error or cluster-robust estimates) and the characteristics of the data. We customize the design of the experiments so that the simulations produce artificial datasets that resemble actual data from studies funded by the International Initiative for Impact Evaluation (3ie). After determining that the SE-ES method performs adequately, we then apply it to the 3ie studies.

We find an average ex post power of 75.4%, not far from the 80% commonly claimed in the 3ie funding applications. However, we find more estimates of low power than would be expected if all studies had 80% true power. Investigation using regression

[8] Kaestner (2021) is another example of low power/large, significant estimate that causes the author to lose confidence in the estimated effect.

analysis reveals that most of the differences between ex post and ex ante power (58%) can be explained by differences in planned and actual total observations, number of clusters, and the degree of intracluster correlation.

A particularly useful application of ex post power estimation is that it can aid in the interpretation of both insignificant and significant estimates. We provide three examples from the 3ie studies: (i) insignificance with low power, (ii) insignificance with high power, and (iii) significance with low power. The first two examples illustrate how one can use the associated power estimate to help determine if statistical insignificance is caused by a negligible effect size or insufficient power. The third example illustrates how ex post power can also be useful when the estimated effect is significant because it can alert the reader to the possibility of Type M error (Gelman & Carlin, 2014).

Limitations. As demonstrated in TABLES 1 and 2, the performance of the SE-ES method depends crucially on the nature of the data and the type of estimator used for estimation. One must be careful in applying the results of the Monte Carlo experiments to settings that are different from those in the experiments. For example, the third example was drawn from a 3ie-funded study in which OLS estimation was applied to a linear equation with a dependent variable taking integer values 1 to 4. Strictly speaking, OLS is not the appropriate estimator to apply in this situation because the dependent variable is bounded by 1 and 4. It is unclear how well heteroskedasticity-robust standard errors accommodate this feature of the data. Accordingly, it is unclear how well the simulation results in TABLE 2 extend to this study. When considering data environments and estimators different from those studied here, one should appropriately customize the Monte Carlo simulations to the data at hand. To facilitate that, the code used to generate

TABLES 1 and 2 (and all the other results in this paper) are publicly available and posted online.[9]

Possible directions for future research. One possible direction for future research is to apply the SE-ES method to funding applications from other organizations to see whether the results we find for 3ie are externally valid. Another direction for future research is to assess the SE-ES method for other types of datasets and estimators. As we demonstrate with TABLES 1 and 2, performance can vary greatly depending on the nature of the data and the properties of the estimators being used. Would the SE-ES method be sufficiently precise to be useful when applied to IV estimation, structural equation models (SEM), or time series applications?

Finally, recent research indicates that much social science research is severely underpowered. Ioannidis, Stanley, & Doucouliagos (2017) report that the median statistical power in economics research is 18%, Using a similar methodology, Arel-Bundock et al. (2022) find that median power in the empirical political science literature is approximately 10%. Both methods use meta-analysis to fix the effect size for their ex post power calculations. An alternative approach would be to work within homogeneous literatures for which one could posit meaningful effect sizes. Then use the methods of this paper to determine whether the respective literatures have sufficient power to identify those effect sizes. It would be of interest to know whether this alternative approach yielded the same conclusions about power as the meta-analytic studies. It is hoped that the present study will stimulate research in these, and other, directions.

[9] See here: https://osf.io/frwx2/?view_only=5a0a8d2ecc2e4f6eb3be8097152f6712.

## References


3ie(a). (n.d.) Push-button replication. Retrieved August 27, 2022 from https://www.3ieimpact.org/our-expertise/replication/push-button-replication.

3ie(b). (n.d.). Transparency, reproducibility, and ethics (TRE) policy: February 2022 (Version 3). Retrieved September 12, 2022 from https://www.3ieimpact.org/sites/default/files/2022-04/3ie-transparent-reproducible-ethical-evidence-policy-2022.pdf

3ie(c). (n.d.). Impact evaluations. Retrieved September 12, 2022 from https://www.3ieimpact.org/evidence-hub/publications/impact-evaluations

Aker, J. C., Boumnijel, R., McClelland, A., & Tierney, N. (2016). Payment mechanisms and antipoverty programs: Evidence from a mobile money cash transfer experiment in Niger. *Economic Development and Cultural Change*, 65(1), 1-37.

Aladysheva, A., Brück, T., Esenaliev, D., Karabaeva, J., Leung, W., & Nillesen, E. (2017). Impact evaluation of the Livingsidebyside peacebuilding educational programme in Kyrgyzstan. *3ie Grantee Final Report.* New Delhi: International Initiative for Impact Evaluation (3ie).

Arel-Bundock, V., Briggs, R., Doucouliagos, H., Aviña, M. M., & Stanley, T. D. (2022). Quantitative political science research is greatly nderpowered. Working paper. https://osf.io/preprints/7vy2f/.

Ashraf, N., Giné, X., & Karlan, D. (2009). Finding missing markets (and a disturbing epilogue): Evidence from an export crop adoption and marketing intervention in Kenya. *American Journal of Agricultural Economics*, 91(4), 973-990.

Bailey, R. C., Moses, S., Parker, C. B., Agot, K., Maclean, I., Krieger, J. N., ... & Ndinya-Achola, J. O. (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The Lancet*, 369(9562), 643-656.

Bellamare, M. (2021, June 30). Top 5 Agricultural Economics Journals–2021 Edition (Updated). Marc F. Bellemare. http://marcfbellemare.com/wordpress/13856

Brown, J. P., Lambert, D. M., & Wojan, T. R. (2019). The effect of the conservation reserve program on rural economies: deriving a statistical verdict from a null finding. *American Journal of Agricultural Economics*, 101(2), 528-540.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414-427.

Center for Open Science, 2022. Non-HSR project definitions. https://osf.io/upywe?view_only=495a1c72f0df4ccd9492962ae38d65e4

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920-80.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Coppock, A. (November 20, 2013). 10 things to know about statistical power. EGAP: Featured Resources. https://egap.org/resource/10-things-to-know-about-statistical-power/

Dercon, S., Gilligan, D. O., Hoddinott, J., & Woldehanna, T. (2009). The impact of agricultural extension and roads on poverty and consumption growth in fifteen Ethiopian villages. *American Journal of Agricultural Economics*, 91(4), 1007-1021.

Djimeu, E. W. & Houndolo, D. G. (2016). Power calculation for causal inference in social science: sample size and minimum detectable effect determination. *Journal of Development Effectiveness*, 8(4), 508-527.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. 38. Society of Industrial and Applied Mathematics CBMS-NSF Monographs. ISBN 0-89871-179-7.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Galiani, S. & Schargrodsky, E. (2010). Property rights for the poor: Effects of land titling. *Journal of Public Economics*, 94(9-10), 700-729.

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. World Bank Publications.

Glennerster, R. & Takavarasha, K. (2013). *Running randomized evaluations*. Princeton University Press.

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19-24.

Huber, C. (10 January 2019). Calculating power using Monte Carlo simulations, part 1: The basics. The Stata Blog. https://blog.stata.com/2019/01/10/calculating-power-using-monte-carlo-simulations-part-1-the-basics/

International Initiative for Impact Evaluation. (2022). Replication studies. https://www.3ieimpact.org/evidence-hub/replication-studies-status

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, 127 (October), F236–F265. doi: 10.1111/ecoj.12461

Kaestner, R. (2021). Mortality and science: A comment on two articles on the effects of health insurance on mortality. *Econ Journal Watch*, 18(2), 192.

Kleinman, K., & Huang, S. S. (2017). Calculating power by bootstrap, with an application to cluster-randomized trials. EGEMs, (Generating Evidence & Methods to improve patient outcomes). 4(1):1-18. DOI: http://doi.org/10.13063/2327-9214.1202

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics/Revue Canadienne d'économique*, 52(3), 851-881.

McKenzie, D. & Ozier, O. (16 May 2019). Why ex-post power using estimated effect sizes is bad, but an ex-post MDE is not. Development Blog. https://blogs.worldbank.org/impactevaluations/why-ex-post-power-using-estimated-effect-sizes-bad-ex-post-mde-not

Miguel, E., & Kremer, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159-217.

Reinikka, R., & Svensson, J. (2005). Fighting corruption to improve schooling: Evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, 3(2-3), 259-267.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal*, 19(1), 4-60.

Skiba, P. M., & Tobacman, J. (2019). Do payday loans cause bankruptcy?. *Journal of Law and Economics,* 62(3), 485-519.

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, 36(10), 1580-1598.

StataCorp. 2021. *Stata 17. Power, Precision, and Sample-Size Reference Manual*. College Station, TX: Stata Press.

Sullivan, P., Hellerstein, D., Hansen, L., Johansson, R., Koenig, S., Lubowski, R. N., ... & Bucholz, S. (2004). The conservation reserve program: economic implications for rural America. *USDA-ERS Agricultural Economic Report*, (834).

Tian, J. (2021). A replication of "The effect of the conservation reserve program on rural economies: Deriving a statistical verdict from a null finding" (American Journal of Agricultural Economics, 2019). Working Paper No. 12/2021, Department of Economics and Finance, University of Canterbury Business School. https://repec.canterbury.ac.nz/cbt/econwp/2112.pdf

Wicklin, R. (30 May 2013). Using simulation to estimate the power of a statistical test. SAS Blogs. https://blogs.sas.com/content/iml/2013/05/30/simulation-power.html

Wicklin, R. (29 October 2018). Bootstrap regression estimates: Residual resampling. SAS Blogs. https://blogs.sas.com/content/iml/2018/10/29/bootstrap-regression-residual-resampling.html

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141-167.

**TABLE 1**
**Performance Assessment of the SE-ES Ex-Post Power Estimator (OLS)**

| *Sample Size* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
|---|---|---|---|---|---|
| | | *True Power = 10%* | | | |
| 3,000 | 0 | 0.100 | 0.098 | 0.103 | 0.002 |
| 10,500 | 0 | 0.100 | 0.099 | 0.101 | 0.001 |
| | | *True Power = 20%* | | | |
| 3,000 | 0 | 0.200 | 0.194 | 0.207 | 0.004 |
| 10,500 | 0 | 0.200 | 0.196 | 0.204 | 0.002 |
| | | *True Power = 30%* | | | |
| 3,000 | 0 | 0.301 | 0.290 | 0.311 | 0.006 |
| 10,500 | 0 | 0.300 | 0.294 | 0.306 | 0.004 |
| | | *True Power = 40%* | | | |
| 3,000 | 0 | 0.400 | 0.387 | 0.414 | 0.008 |
| 10,500 | 0 | 0.400 | 0.393 | 0.407 | 0.005 |
| | | *True Power = 50%* | | | |
| 3,000 | 0 | 0.501 | 0.484 | 0.520 | 0.011 |
| 10,500 | 0 | 0.500 | 0.491 | 0.509 | 0.005 |
| | | *True Power = 60%* | | | |
| 3,000 | 0 | 0.601 | 0.583 | 0.619 | 0.011 |
| 10,500 | 0 | 0.600 | 0.590 | 0.609 | 0.006 |
| | | *True Power = 70%* | | | |
| 3,000 | 0 | 0.700 | 0.681 | 0.719 | 0.011 |
| 10,500 | 0 | 0.700 | 0.690 | 0.710 | 0.006 |
| | | *True Power = 80%* | | | |
| 3,000 | 0 | 0.801 | 0.783 | 0.817 | 0.010 |
| 10,500 | 0 | 0.800 | 0.791 | 0.809 | 0.005 |
| | | *True Power = 90%* | | | |
| 3,000 | 0 | 0.900 | 0.888 | 0.911 | 0.007 |
| 10,500 | 0 | 0.900 | 0.894 | 0.906 | 0.004 |

NOTE: The Monte Carlo experiments that produced these results are described in Section IV. The experiments for sample size $= 3{,}000$ used $N \times T = 60 \times 50$, and the experiments for sample size $= 10{,}500$ used $N \times T = 150 \times 70$. Both sets of experiments set $\rho = 0$. Each experiment consisted of 1000 replications.

TABLE 2
Performance Assessment of the SE-ES Ex-Post Power Estimator (OLS-Cluster)

| *True Power = 10%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.103 | 0.086 | 0.125 | 0.012 |
| (60, 50) | 0.150 | 0.103 | 0.087 | 0.126 | 0.012 |
| (60, 50) | 0.250 | 0.103 | 0.087 | 0.124 | 0.012 |
| (100, 30) | 0.050 | 0.102 | 0.089 | 0.117 | 0.009 |
| (100, 30) | 0.150 | 0.102 | 0.089 | 0.118 | 0.009 |
| (100, 30) | 0.250 | 0.102 | 0.088 | 0.118 | 0.009 |
| (150, 70) | 0.050 | 0.101 | 0.090 | 0.114 | 0.007 |
| (150, 70) | 0.150 | 0.101 | 0.090 | 0.114 | 0.007 |
| (150, 70) | 0.250 | 0.101 | 0.090 | 0.114 | 0.007 |
| (250, 42) | 0.050 | 0.100 | 0.092 | 0.110 | 0.005 |
| (250, 42) | 0.150 | 0.101 | 0.092 | 0.109 | 0.005 |
| (250, 42) | 0.250 | 0.101 | 0.092 | 0.110 | 0.005 |
| *True Power = 20%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.208 | 0.164 | 0.265 | 0.033 |
| (60, 50) | 0.150 | 0.208 | 0.164 | 0.268 | 0.032 |
| (60, 50) | 0.250 | 0.208 | 0.163 | 0.267 | 0.032 |
| (100, 30) | 0.050 | 0.206 | 0.171 | 0.251 | 0.026 |
| (100, 30) | 0.150 | 0.206 | 0.170 | 0.251 | 0.026 |
| (100, 30) | 0.250 | 0.206 | 0.170 | 0.251 | 0.025 |
| (150, 70) | 0.050 | 0.204 | 0.175 | 0.237 | 0.019 |
| (150, 70) | 0.150 | 0.204 | 0.175 | 0.235 | 0.019 |
| (150, 70) | 0.250 | 0.204 | 0.174 | 0.236 | 0.019 |
| (250, 42) | 0.050 | 0.202 | 0.180 | 0.228 | 0.015 |
| (250, 42) | 0.150 | 0.202 | 0.180 | 0.227 | 0.015 |
| (250, 42) | 0.250 | 0.202 | 0.180 | 0.227 | 0.015 |

| *True Power = 30%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.314 | 0.243 | 0.410 | 0.052 |
| (60, 50) | 0.150 | 0.314 | 0.239 | 0.408 | 0.053 |
| (60, 50) | 0.250 | 0.315 | 0.238 | 0.411 | 0.052 |
| (100, 30) | 0.050 | 0.309 | 0.253 | 0.375 | 0.038 |
| (100, 30) | 0.150 | 0.310 | 0.254 | 0.377 | 0.038 |
| (100, 30) | 0.250 | 0.310 | 0.254 | 0.377 | 0.039 |
| (150, 70) | 0.050 | 0.304 | 0.259 | 0.356 | 0.029 |
| (150, 70) | 0.150 | 0.304 | 0.259 | 0.355 | 0.029 |
| (150, 70) | 0.250 | 0.304 | 0.259 | 0.355 | 0.029 |
| (250, 42) | 0.050 | 0.304 | 0.268 | 0.342 | 0.023 |
| (250, 42) | 0.150 | 0.303 | 0.268 | 0.342 | 0.023 |
| (250, 42) | 0.250 | 0.303 | 0.268 | 0.343 | 0.023 |
| *True Power = 40%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.415 | 0.319 | 0.534 | 0.066 |
| (60, 50) | 0.150 | 0.413 | 0.317 | 0.530 | 0.065 |
| (60, 50) | 0.250 | 0.412 | 0.317 | 0.525 | 0.065 |
| (100, 30) | 0.050 | 0.408 | 0.333 | 0.495 | 0.049 |
| (100, 30) | 0.150 | 0.408 | 0.330 | 0.499 | 0.050 |
| (100, 30) | 0.250 | 0.408 | 0.331 | 0.501 | 0.051 |
| (150, 70) | 0.050 | 0.404 | 0.344 | 0.476 | 0.041 |
| (150, 70) | 0.150 | 0.404 | 0.344 | 0.475 | 0.041 |
| (150, 70) | 0.250 | 0.404 | 0.344 | 0.475 | 0.041 |
| (250, 42) | 0.050 | 0.402 | 0.356 | 0.456 | 0.030 |
| (250, 42) | 0.150 | 0.403 | 0.356 | 0.454 | 0.030 |
| (250, 42) | 0.250 | 0.402 | 0.358 | 0.456 | 0.031 |

| *True Power = 50%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.518 | 0.404 | 0.652 | 0.076 |
| (60, 50) | 0.150 | 0.517 | 0.399 | 0.650 | 0.076 |
| (60, 50) | 0.250 | 0.517 | 0.396 | 0.646 | 0.075 |
| (100, 30) | 0.050 | 0.510 | 0.428 | 0.602 | 0.055 |
| (100, 30) | 0.150 | 0.510 | 0.423 | 0.606 | 0.055 |
| (100, 30) | 0.250 | 0.510 | 0.422 | 0.604 | 0.055 |
| (150, 70) | 0.050 | 0.508 | 0.440 | 0.585 | 0.046 |
| (150, 70) | 0.150 | 0.508 | 0.437 | 0.584 | 0.046 |
| (150, 70) | 0.250 | 0.508 | 0.436 | 0.587 | 0.046 |
| (250, 42) | 0.050 | 0.505 | 0.450 | 0.564 | 0.035 |
| (250, 42) | 0.150 | 0.505 | 0.453 | 0.567 | 0.035 |
| (250, 42) | 0.250 | 0.506 | 0.451 | 0.566 | 0.035 |
| *True Power = 60%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.613 | 0.481 | 0.758 | 0.083 |
| (60, 50) | 0.150 | 0.614 | 0.489 | 0.762 | 0.082 |
| (60, 50) | 0.250 | 0.615 | 0.487 | 0.754 | 0.082 |
| (100, 30) | 0.050 | 0.606 | 0.510 | 0.718 | 0.062 |
| (100, 30) | 0.150 | 0.606 | 0.509 | 0.712 | 0.061 |
| (100, 30) | 0.250 | 0.606 | 0.510 | 0.710 | 0.060 |
| (150, 70) | 0.050 | 0.606 | 0.527 | 0.692 | 0.050 |
| (150, 70) | 0.150 | 0.606 | 0.526 | 0.689 | 0.050 |
| (150, 70) | 0.250 | 0.606 | 0.529 | 0.688 | 0.049 |
| (250, 42) | 0.050 | 0.604 | 0.546 | 0.667 | 0.037 |
| (250, 42) | 0.150 | 0.604 | 0.545 | 0.667 | 0.037 |
| (250, 42) | 0.250 | 0.604 | 0.543 | 0.669 | 0.038 |

| *True Power = 70%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.711 | 0.581 | 0.841 | 0.080 |
| (60, 50) | 0.150 | 0.711 | 0.578 | 0.843 | 0.079 |
| (60, 50) | 0.250 | 0.711 | 0.581 | 0.833 | 0.078 |
| (100, 30) | 0.050 | 0.707 | 0.608 | 0.812 | 0.063 |
| (100, 30) | 0.150 | 0.708 | 0.607 | 0.811 | 0.062 |
| (100, 30) | 0.250 | 0.708 | 0.608 | 0.810 | 0.061 |
| (150, 70) | 0.050 | 0.707 | 0.624 | 0.790 | 0.050 |
| (150, 70) | 0.150 | 0.707 | 0.625 | 0.793 | 0.051 |
| (150, 70) | 0.250 | 0.707 | 0.627 | 0.795 | 0.051 |
| (250, 42) | 0.050 | 0.705 | 0.643 | 0.767 | 0.039 |
| (250, 42) | 0.150 | 0.705 | 0.642 | 0.768 | 0.039 |
| (250, 42) | 0.250 | 0.705 | 0.642 | 0.768 | 0.039 |
| *True Power = 80%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.806 | 0.680 | 0.915 | 0.070 |
| (60, 50) | 0.150 | 0.806 | 0.685 | 0.918 | 0.071 |
| (60, 50) | 0.250 | 0.806 | 0.689 | 0.917 | 0.070 |
| (100, 30) | 0.050 | 0.807 | 0.718 | 0.891 | 0.052 |
| (100, 30) | 0.150 | 0.806 | 0.714 | 0.891 | 0.053 |
| (100, 30) | 0.250 | 0.806 | 0.713 | 0.890 | 0.053 |
| (150, 70) | 0.050 | 0.801 | 0.726 | 0.874 | 0.045 |
| (150, 70) | 0.150 | 0.802 | 0.728 | 0.873 | 0.044 |
| (150, 70) | 0.250 | 0.802 | 0.728 | 0.868 | 0.044 |
| (250, 42) | 0.050 | 0.801 | 0.741 | 0.856 | 0.035 |
| (250, 42) | 0.150 | 0.802 | 0.746 | 0.857 | 0.035 |
| (250, 42) | 0.250 | 0.802 | 0.746 | 0.856 | 0.035 |

| *True Power = 90%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.903 | 0.806 | 0.973 | 0.051 |
| (60, 50) | 0.150 | 0.902 | 0.810 | 0.972 | 0.050 |
| (60, 50) | 0.250 | 0.901 | 0.809 | 0.972 | 0.050 |
| (100, 30) | 0.050 | 0.902 | 0.829 | 0.959 | 0.039 |
| (100, 30) | 0.150 | 0.902 | 0.832 | 0.959 | 0.039 |
| (100, 30) | 0.250 | 0.902 | 0.834 | 0.958 | 0.039 |
| (150, 70) | 0.050 | 0.900 | 0.843 | 0.952 | 0.034 |
| (150, 70) | 0.150 | 0.901 | 0.841 | 0.952 | 0.033 |
| (150, 70) | 0.250 | 0.901 | 0.843 | 0.951 | 0.033 |
| (250, 42) | 0.050 | 0.901 | 0.854 | 0.940 | 0.026 |
| (250, 42) | 0.150 | 0.901 | 0.857 | 0.940 | 0.026 |
| (250, 42) | 0.250 | 0.901 | 0.858 | 0.939 | 0.025 |

NOTE: The Monte Carlo experiments that produced these results are described in Section IV. Each experiment consisted of 1000 replications.

**TABLE 3**
**Comparison of Ex Post with Ex Ante Power and Sample Characteristics**

| ID | Power | | Observations | | Clusters | | ICC | |
|---|---|---|---|---|---|---|---|---|
| | Ex Ante | Ex Post | Ex Ante | Ex Post | Ex Ante | Ex Post | Ex Ante | Ex Post |
| 1 | 80.0% | 84.0% | 2,520 | 2,991 | 402 | 393 | 0.030 | 0.085 |
| 1 | 80.0% | 86.3% | 2,520 | 3,431 | 402 | 393 | 0.030 | 0.205 |
| 2 | 80.0% | 100.0% | 3,240 | 3,128 | 108 | 107 | 0.500 | 0.070 |
| 2 | 80.0% | 100.0% | 3,240 | 2,853 | 108 | 107 | 0.500 | 0.060 |
| 3 | 80.0% | 75.0% | 12,000 | 11,733 | 300 | 301 | n.a | 0.300 |
| 3 | 80.0% | 76.8% | 12,000 | 11,733 | 300 | 301 | n.a. | 0.250 |
| 4 | 80.0% | 31.9% | 2,808 | 2,834 | 236 | 234 | n.a. | 0.160 |
| 4 | 80.0% | 70.2% | 2,808 | 2,839 | 236 | 234 | n.a. | 0.010 |
| 5 | 80.0% | 69.8% | 7,200 | 6,085 | 80 | 120 | 0.031 | 0.144 |
| 5 | 80.0% | 78.1% | 7,200 | 6,015 | 80 | 120 | 0.030 | 0.115 |
| 6 | 80.0% | 92.9% | 4,606 | 4,158 | 102 | 101 | 0.023 | 0.145 |
| 6 | 80.0% | 99.5% | 4,606 | 4,156 | 102 | 101 | 0.023 | 0.145 |
| 7 | 80.0% | 69.6% | 2,717 | 2,483 | 90 | 94 | 0.070 | 0.165 |
| 8 | 80.0% | 51.8% | 2,160 | 2,131 | 80 | 48 | 0.200 | 0.310 |
| 8 | 80.0% | 66.2% | 2,160 | 2,131 | 80 | 48 | 0.200 | 0.350 |
| 8 | 80.0% | 75.1% | 2,160 | 2,131 | 80 | 48 | 0.200 | 0.370 |
| 9 | 80.0% | 99.8% | 16,880 | 16,827 | 62 | 66 | 0.103 | 0.050 |
| 10 | 80.0% | 68.0% | 2,169 | 2,687 | 120 | 110 | 0.202 | 0.268 |
| 10 | 80.0% | 85.1% | 2,169 | 2,679 | 120 | 110 | 0.202 | 0.255 |
| 11 | 80.0% | 63.8% | 4,009 | 2,358 | n.a. | 62 | n.a. | 0.020 |
| 11 | 80.0% | 70.6% | 4,009 | 2,358 | n.a. | 62 | n.a. | 0.050 |
| 12 | 80.0% | 87.4% | 1,948 | 2,694 | 60 | 68 | 0.100 | 0.017 |
| 12 | 80.0% | 94.4% | 1,948 | 2,694 | 60 | 68 | 0.100 | 0.044 |
| 12 | 80.0% | 100.0% | 1,948 | 2,694 | 60 | 68 | 0.100 | 0.026 |

| ID | Power | | Observations | | Clusters | | ICC | |
|---|---|---|---|---|---|---|---|---|
| | Ex Ante | Ex Post | Ex Ante | Ex Post | Ex Ante | Ex Post | Ex Ante | Ex Post |
| 13 | 80.0% | 20.7% | 10,070 | 9,897 | 118 | 63 | 0.070 | 0.098 |
| 14 | 80.0% | 90.3% | 4,378 | 4,017 | 300 | 246 | n.a. | 0.100 |
| 14 | 80.0% | 96.6% | 4,378 | 4,191 | 300 | 246 | n.a. | 0.150 |
| 15 | 80.0% | 86.7% | 2,601 | 2,519 | 80 | 154 | 0.100 | 0.080 |
| 15 | 80.0% | 88.6% | 2,601 | 2,531 | 80 | 154 | 0.100 | 0.020 |
| 15 | 80.0% | 90.4% | 2,601 | 2,522 | 80 | 154 | 0.100 | 0.160 |
| 15 | 80.0% | 92.1% | 2,601 | 2,525 | 80 | 154 | 0.100 | 0.020 |
| 16 | 80.0% | 90.7% | 2,065 | 1,875 | 100 | 105 | 0.060 | 0.030 |
| 16 | 80.0% | 92.9% | 2,065 | 1,875 | 100 | 106 | 0.060 | 0.030 |
| 16 | 80.0% | 96.8% | 2,065 | 1,875 | 100 | 105 | 0.060 | 0.030 |
| 17 | 80.0% | 43.9% | 22,578 | 14,713 | 173 | 122 | 0.150 | 0.359 |
| 17 | 80.0% | 59.1% | 22,578 | 14,713 | 173 | 122 | 0.150 | 0.245 |
| 17 | 80.0% | 63.8% | 22,578 | 14,713 | 173 | 122 | 0.150 | 0.223 |
| 18 | 80.0% | 32.7% | 1,800 | 1,676 | 60 | 20 | 0.250 | 0.140 |
| 18 | 80.0% | 37.0% | 1,800 | 1,316 | 60 | 20 | 0.250 | 0.150 |
| 19 | 90.0% | 71.6% | 10,333 | 12,881 | 120 | 157 | 0.150 | 0.060 |
| 20 | 90.0% | 36.0% | 4,000 | 831 | n.a. | 216 | 0.010 | 0.055 |
| 20 | 90.0% | 100.0% | 4,000 | 3,368 | n.a. | 216 | 0.010 | 0.020 |
| 21 | 80.0% | 68.0% | 3,750 | 3,511 | 300 | 148 | n.a. | 0.180 |
| 22 | 85.0% | 92.6% | 1,858 | 1,798 | 80 | 79 | n.a. | 0.001 |
| 22 | 85.0% | 94.1% | 1,858 | 1,798 | 80 | 80 | n.a. | 0.001 |
| 23 | 80.0% | 42.1% | 5,000 | 6,300 | 100 | 98 | 0.040 | 0.113 |
| 23 | 80.0% | 62.9% | 5,000 | 6,300 | 100 | 98 | 0.040 | 0.057 |
| **Average** | **80.9%** | **75.4%** | **5352.8** | **4756.8** | **140.1** | **135.1** | **0.125** | **0.126** |

SOURCE: All variables other than Ex Post Power come from data supplied by 3ie. The Ex Post Power values are calculated using the SE-ES method described in Section III. Effect sizes were taken from the respective studies ex ante power calculations included as part of their grant applications.

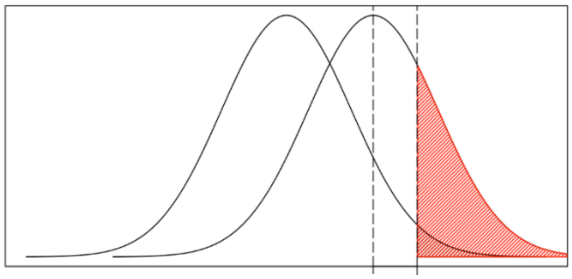
**TABLE 4**
**Determinants of Differences Between Ex Post and Ex Ante Power Estimates**

| Variables | Estimates | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| ΔObs | 0.1375<br>(0.1143) | 0.0067<br>(0.1254) | 0.1046<br>(0.1209) | -0.0550<br>(0.1893) |
| ΔClust | 0.2599***<br>(0.0670) | 0.4401***<br>(0.0726) | 0.1957*<br>(0.0956) | 0.5046***<br>(0.1637) |
| ΔICC | -34.873*<br>(17.414) | 21.530<br>(24.491) | -43.463**<br>(17.188) | -18.588<br>(31.337) |
| ΔObs_sq | ---- | -0.00388<br>(0.00451) | ---- | -0.00482<br>(0.01132) |
| ΔClust_sq | ---- | -0.00444***<br>(0.00096) | ---- | -0.00515***<br>(0.00147) |
| ΔICC_sq | ---- | 54.893<br>(68.951) | ---- | 26.667<br>(75.621) |
| ΔObs × ΔClust | ---- | ---- | -0.00842<br>(0.00622) | 0.00551<br>(0.01520) |
| ΔObs × ΔICC | ---- | ---- | -0.02678<br>(1.19549) | 1.55974<br>(1.05026) |
| ΔClust × ΔICC | ---- | ---- | -0.73456<br>(0.64371) | -0.66459<br>(0.52057) |
| N | 34 | 34 | 34 | 34 |
| $R^2$ | 0.373 | 0.578 | 0.408 | 0.601 |
| BIC | 307.1 | 304.2 | 315.7 | 312.8 |
| Hypothesis Test:<br>Squared Terms = 0 | ---- | F=4.36<br>(p=0.012) | ---- | F=3.89<br>(p=0.021) |
| Hypothesis Test:<br>Interaction Terms = 0 | ---- | ---- | F= 0.53<br>(p=0.664) | F= 0.48<br>(p=0.701) |
| Hypothesis Test:<br>Squared Terms + Interaction Terms = 0 | ---- | ---- | ---- | F= 2.29<br>(p=0.068) |

NOTE: The dependent variable is ΔClust. Estimates are obtained from OLS regressions using robust cluster estimates of standard errors. Data for the regressions come from TABLE 3. Standard errors are reported in parentheses unless p-values are indicated. ***, **, and * indicate statistical significance at the 1 percent, 5 percent, and 10 percent levels.

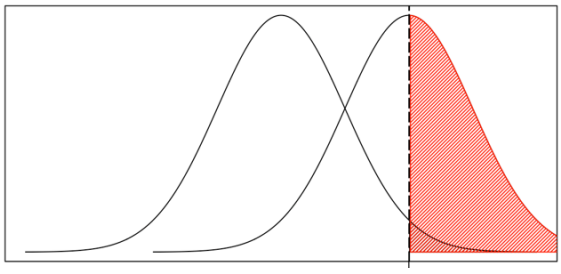
**FIGURE 1**
**The Relationship Between Effect Size and Power**

**A. Case One**: $t_{\upsilon,1-\alpha/2} > t_{ES}$[10]



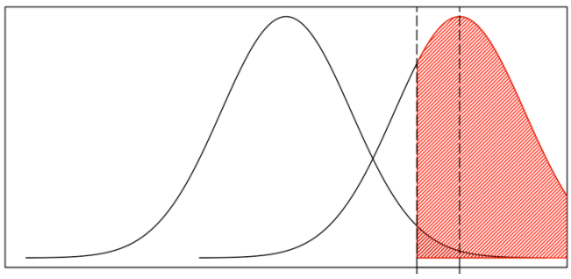

| $t_{\upsilon,1-\alpha/2}$ | $ES$ | $s.e.(\widehat{ES})$ | $t_{ES}$ | $\left(t_{\upsilon,1-\alpha/2} - t_{ES}\right)$ | $1 - Power$ | $Power$ |
|---|---|---|---|---|---|---|
| 2 | 4 | 3 | 1.33 | 0.67 | 0.747 | 0.253 |

**B. Case Two**: $t_{\upsilon,1-\alpha/2} = t_{ES}$




| $t_{\upsilon,1-\alpha/2}$ | $ES$ | $s.e.(\widehat{ES})$ | $t_{ES}$ | $\left(t_{\upsilon,1-\alpha/2} - t_{ES}\right)$ | $1 - Power$ | $Power$ |
|---|---|---|---|---|---|---|
| 2 | 4 | 2 | 2 | 0 | 0.500 | 0.500 |

**C. Case Three**: $t_{\upsilon,1-\alpha/2} < t_{ES}$

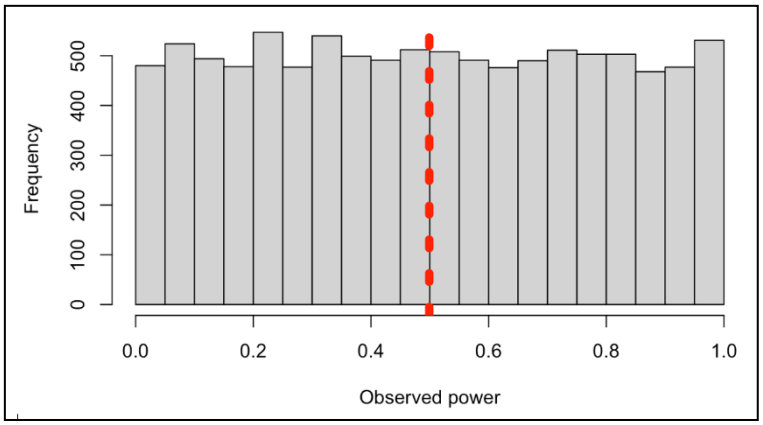



| $t_{\upsilon,1-\alpha/2}$ | $ES$ | $s.e.(\widehat{ES})$ | $t_{ES}$ | $\left(t_{\upsilon,1-\alpha/2} - t_{ES}\right)$ | $1 - Power$ | $Power$ |
|---|---|---|---|---|---|---|
| 2 | 4 | 1.5 | 2.67 | -0.67 | 0.253 | 0.747 |

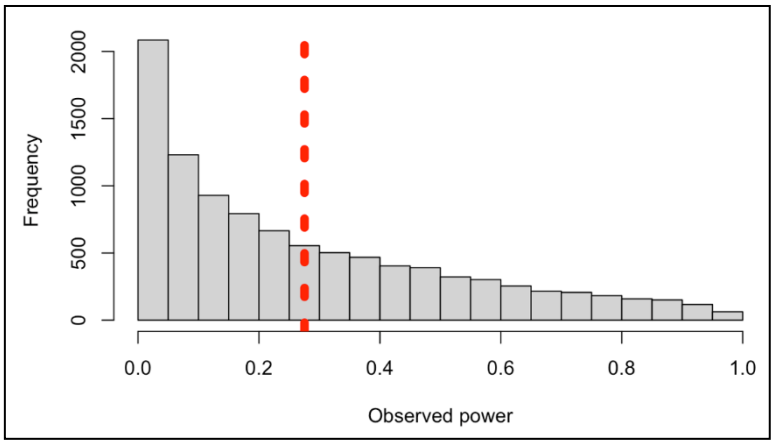
[10] All three cases set $\alpha = 0.05$ and $\upsilon = 50$.

**FIGURE 2**
**Distribution of Observed Power for Different True Power Values**

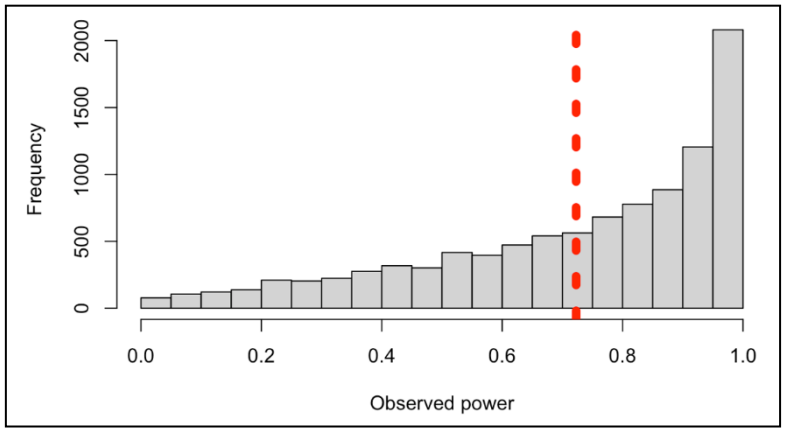
CASE ONE: True Power = 50%




CASE TWO: True Power = 20%




CASE THREE: True Power = 80%



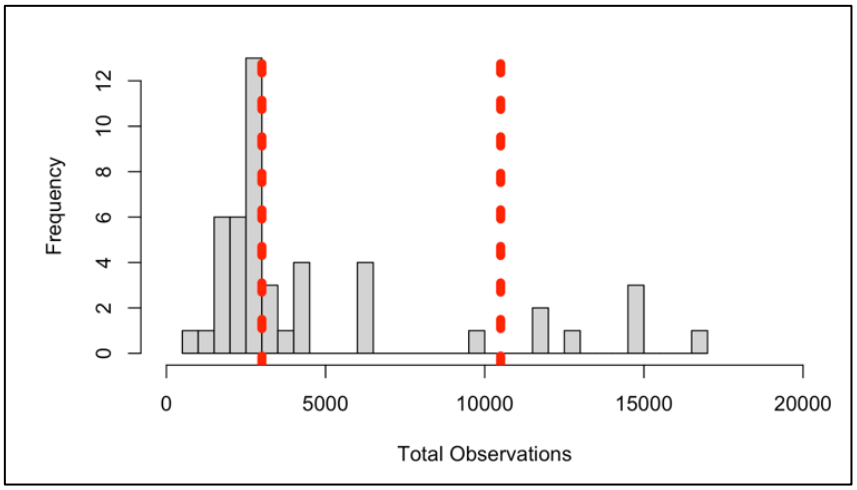
NOTE: The data generating processes used to produce the three histograms above are given below:

- 50%: Y = 1000+ 3.92*Treatment + rnorm(10000, 0, sd=100)
- 20%: Y = 1000+ 2.238*Treatment + rnorm(10000, 0, sd=100)
- 80%: Y = 1000+ 5.6*Treatment + rnorm(10000, 0, sd=100)

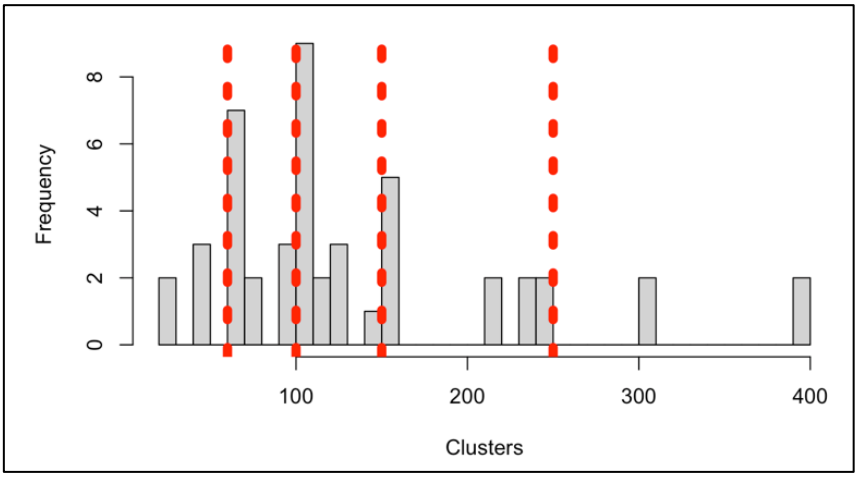
where Treatment is a binary variable consisting of half 1s and half 0s, and rnorm(10000, 0, sd=100) produces a vector of 10,000 realizations from a normal distribution having mean 0 and standard deviation 100.

**FIGURE 3**
**Data Characteristics of the 3ie Samples**

A. Total Observations




B. Clusters




C. ICC

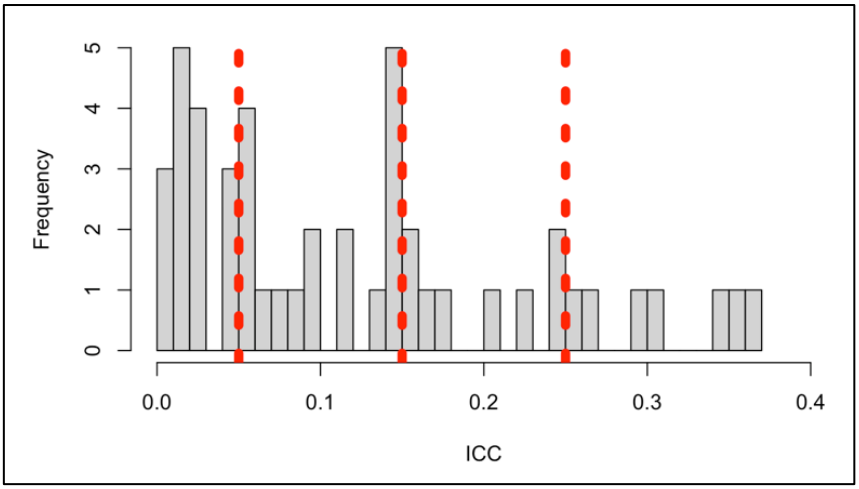
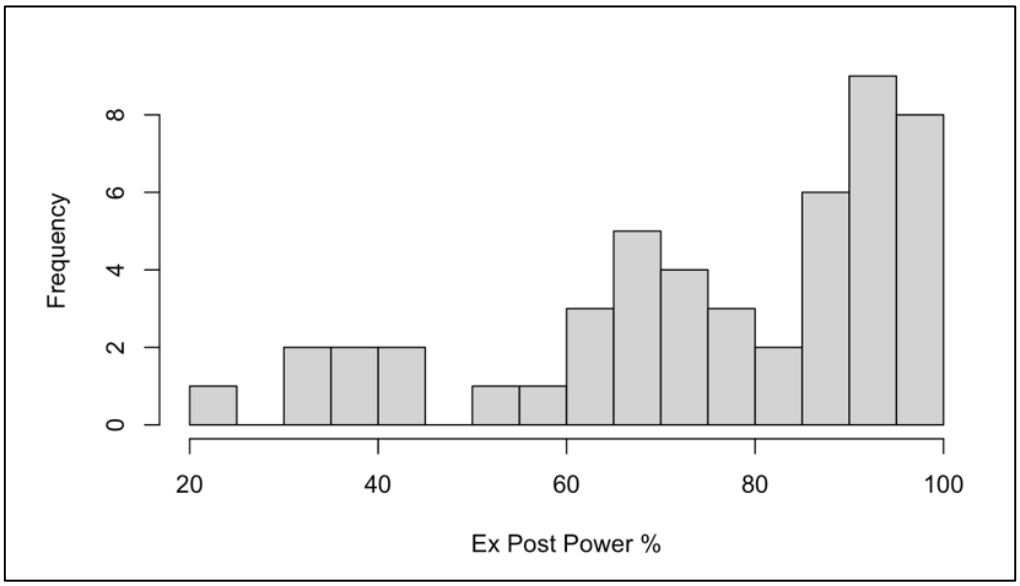


**FIGURE 4**
**Distribution of Ex Post Power Estimates**

**A. Histogram of Ex Post Power Estimates**




**B. Comparison of Ex Ante and Ex Post Power Estimates**

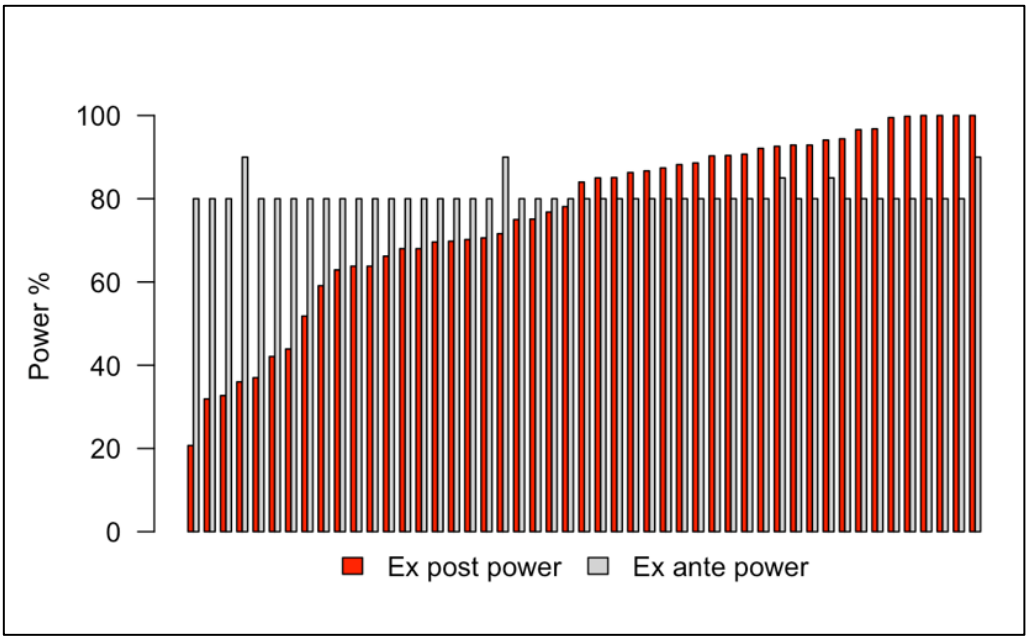
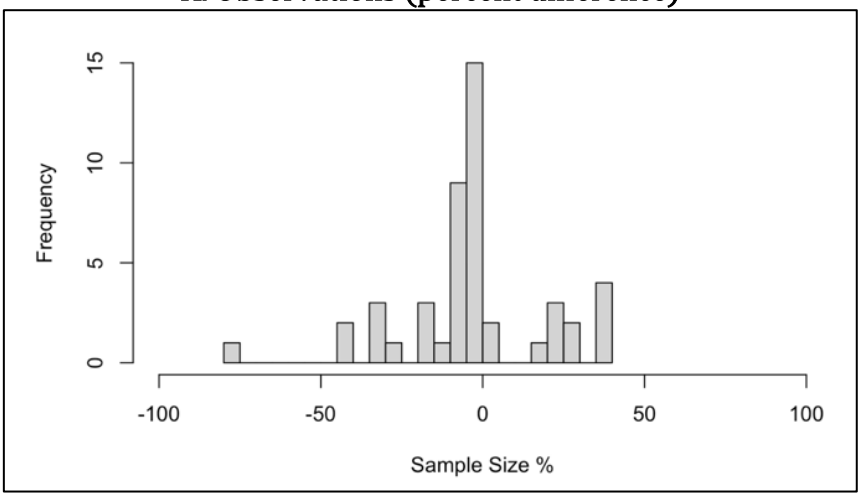
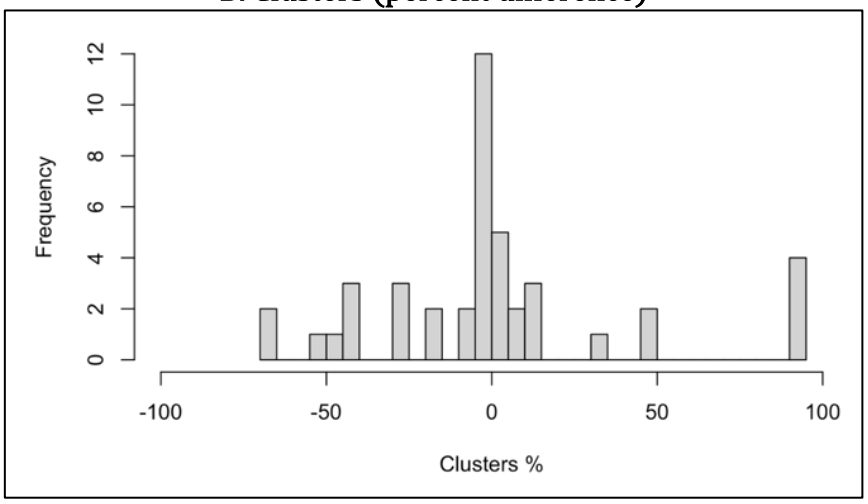


FIGURE 5
Comparison of Planned and Actual Sample Characteristics

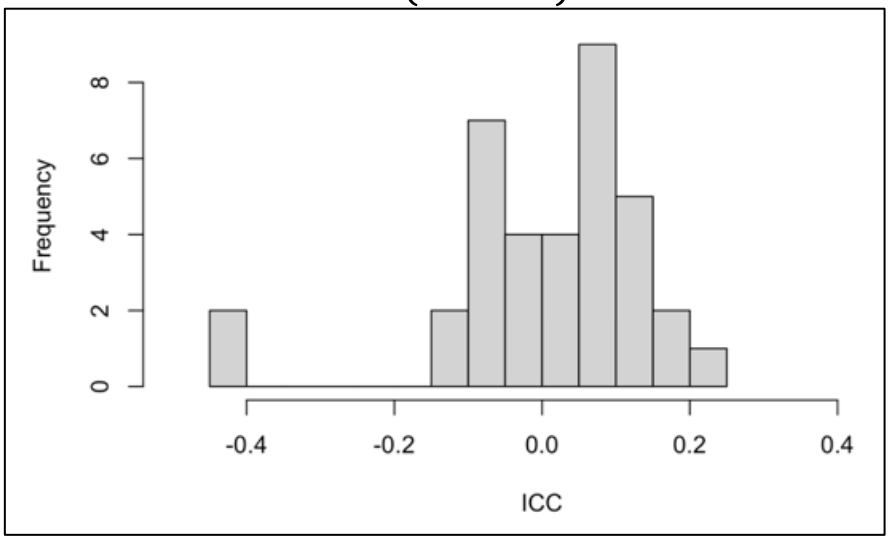
A. Observations (percent difference)




B. Clusters (percent difference)




C. ICC (difference)



## APPENDIX A
## Performance Assessment of the SE-ES Ex-Post Power Estimator (OLS-Cluster) with Cross-sectional Correlation

| *True Power = 10%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.103 | 0.086 | 0.125 | 0.012 |
| (60, 50) | 0.150 | 0.104 | 0.086 | 0.125 | 0.012 |
| (60, 50) | 0.250 | 0.104 | 0.086 | 0.125 | 0.012 |
| (100, 30) | 0.050 | 0.102 | 0.089 | 0.117 | 0.009 |
| (100, 30) | 0.150 | 0.102 | 0.089 | 0.117 | 0.009 |
| (100, 30) | 0.250 | 0.102 | 0.089 | 0.117 | 0.009 |
| (150, 70) | 0.050 | 0.101 | 0.090 | 0.113 | 0.007 |
| (150, 70) | 0.150 | 0.100 | 0.098 | 0.103 | 0.002 |
| (150, 70) | 0.250 | 0.101 | 0.090 | 0.113 | 0.007 |
| (250, 42) | 0.050 | 0.100 | 0.092 | 0.109 | 0.005 |
| (250, 42) | 0.150 | 0.100 | 0.093 | 0.110 | 0.005 |
| (250, 42) | 0.250 | 0.100 | 0.092 | 0.110 | 0.005 |
| *True Power = 20%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.208 | 0.163 | 0.266 | 0.033 |
| (60, 50) | 0.150 | 0.209 | 0.162 | 0.272 | 0.034 |
| (60, 50) | 0.250 | 0.209 | 0.163 | 0.272 | 0.034 |
| (100, 30) | 0.050 | 0.205 | 0.170 | 0.252 | 0.026 |
| (100, 30) | 0.150 | 0.205 | 0.170 | 0.252 | 0.026 |
| (100, 30) | 0.250 | 0.205 | 0.170 | 0.252 | 0.026 |
| (150, 70) | 0.050 | 0.203 | 0.175 | 0.237 | 0.019 |
| (150, 70) | 0.150 | 0.202 | 0.186 | 0.219 | 0.010 |
| (150, 70) | 0.250 | 0.203 | 0.175 | 0.236 | 0.019 |
| (250, 42) | 0.050 | 0.202 | 0.180 | 0.228 | 0.015 |
| (250, 42) | 0.150 | 0.202 | 0.181 | 0.228 | 0.015 |
| (250, 42) | 0.250 | 0.202 | 0.180 | 0.228 | 0.015 |

| *True Power = 30%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.314 | 0.242 | 0.410 | 0.052 |
| (60, 50) | 0.150 | 0.314 | 0.242 | 0.406 | 0.051 |
| (60, 50) | 0.250 | 0.314 | 0.242 | 0.407 | 0.051 |
| (100, 30) | 0.050 | 0.308 | 0.254 | 0.375 | 0.038 |
| (100, 30) | 0.150 | 0.308 | 0.254 | 0.375 | 0.038 |
| (100, 30) | 0.250 | 0.308 | 0.254 | 0.375 | 0.038 |
| (150, 70) | 0.050 | 0.306 | 0.259 | 0.358 | 0.030 |
| (150, 70) | 0.150 | 0.303 | 0.274 | 0.336 | 0.019 |
| (150, 70) | 0.250 | 0.305 | 0.259 | 0.359 | 0.030 |
| (250, 42) | 0.050 | 0.303 | 0.269 | 0.342 | 0.023 |
| (250, 42) | 0.150 | 0.303 | 0.268 | 0.341 | 0.023 |
| (250, 42) | 0.250 | 0.303 | 0.268 | 0.341 | 0.023 |
| *True Power = 40%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.416 | 0.319 | 0.534 | 0.067 |
| (60, 50) | 0.150 | 0.416 | 0.318 | 0.536 | 0.067 |
| (60, 50) | 0.250 | 0.416 | 0.317 | 0.535 | 0.067 |
| (100, 30) | 0.050 | 0.409 | 0.335 | 0.494 | 0.048 |
| (100, 30) | 0.150 | 0.409 | 0.334 | 0.495 | 0.048 |
| (100, 30) | 0.250 | 0.409 | 0.334 | 0.496 | 0.048 |
| (150, 70) | 0.050 | 0.405 | 0.347 | 0.479 | 0.040 |
| (150, 70) | 0.150 | 0.403 | 0.363 | 0.455 | 0.028 |
| (150, 70) | 0.250 | 0.405 | 0.347 | 0.480 | 0.040 |
| (250, 42) | 0.050 | 0.402 | 0.355 | 0.455 | 0.030 |
| (250, 42) | 0.150 | 0.402 | 0.355 | 0.455 | 0.030 |
| (250, 42) | 0.250 | 0.402 | 0.355 | 0.455 | 0.030 |

| *True Power = 50%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.516 | 0.402 | 0.644 | 0.074 |
| (60, 50) | 0.150 | 0.516 | 0.403 | 0.644 | 0.074 |
| (60, 50) | 0.250 | 0.516 | 0.404 | 0.646 | 0.074 |
| (100, 30) | 0.050 | 0.510 | 0.423 | 0.603 | 0.055 |
| (100, 30) | 0.150 | 0.510 | 0.423 | 0.602 | 0.055 |
| (100, 30) | 0.250 | 0.510 | 0.422 | 0.603 | 0.055 |
| (150, 70) | 0.050 | 0.508 | 0.437 | 0.586 | 0.046 |
| (150, 70) | 0.150 | 0.506 | 0.454 | 0.563 | 0.034 |
| (150, 70) | 0.250 | 0.508 | 0.438 | 0.585 | 0.046 |
| (250, 42) | 0.050 | 0.503 | 0.449 | 0.561 | 0.035 |
| (250, 42) | 0.150 | 0.503 | 0.449 | 0.562 | 0.035 |
| (250, 42) | 0.250 | 0.503 | 0.449 | 0.562 | 0.035 |
| *True Power = 60%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.612 | 0.478 | 0.758 | 0.083 |
| (60, 50) | 0.150 | 0.612 | 0.479 | 0.758 | 0.083 |
| (60, 50) | 0.250 | 0.612 | 0.480 | 0.758 | 0.083 |
| (100, 30) | 0.050 | 0.607 | 0.513 | 0.715 | 0.061 |
| (100, 30) | 0.150 | 0.607 | 0.512 | 0.715 | 0.061 |
| (100, 30) | 0.250 | 0.607 | 0.512 | 0.716 | 0.061 |
| (150, 70) | 0.050 | 0.606 | 0.525 | 0.693 | 0.050 |
| (150, 70) | 0.150 | 0.605 | 0.543 | 0.672 | 0.038 |
| (150, 70) | 0.250 | 0.606 | 0.526 | 0.693 | 0.050 |
| (250, 42) | 0.050 | 0.603 | 0.547 | 0.667 | 0.037 |
| (250, 42) | 0.150 | 0.603 | 0.547 | 0.666 | 0.037 |
| (250, 42) | 0.250 | 0.603 | 0.547 | 0.666 | 0.037 |

| *True Power = 70%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.712 | 0.583 | 0.842 | 0.079 |
| (60, 50) | 0.150 | 0.712 | 0.584 | 0.842 | 0.079 |
| (60, 50) | 0.250 | 0.712 | 0.584 | 0.843 | 0.079 |
| (100, 30) | 0.050 | 0.705 | 0.601 | 0.811 | 0.063 |
| (100, 30) | 0.150 | 0.705 | 0.601 | 0.809 | 0.063 |
| (100, 30) | 0.250 | 0.706 | 0.602 | 0.808 | 0.063 |
| (150, 70) | 0.050 | 0.707 | 0.626 | 0.786 | 0.049 |
| (150, 70) | 0.150 | 0.706 | 0.643 | 0.769 | 0.039 |
| (150, 70) | 0.250 | 0.706 | 0.627 | 0.787 | 0.049 |
| (250, 42) | 0.050 | 0.705 | 0.642 | 0.769 | 0.039 |
| (250, 42) | 0.150 | 0.705 | 0.641 | 0.769 | 0.039 |
| (250, 42) | 0.250 | 0.705 | 0.641 | 0.769 | 0.039 |
| *True Power = 80%* | | | | | |
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.805 | 0.683 | 0.912 | 0.069 |
| (60, 50) | 0.150 | 0.805 | 0.683 | 0.911 | 0.069 |
| (60, 50) | 0.250 | 0.805 | 0.682 | 0.911 | 0.069 |
| (100, 30) | 0.050 | 0.806 | 0.717 | 0.891 | 0.053 |
| (100, 30) | 0.150 | 0.807 | 0.716 | 0.891 | 0.053 |
| (100, 30) | 0.250 | 0.807 | 0.715 | 0.891 | 0.053 |
| (150, 70) | 0.050 | 0.801 | 0.724 | 0.874 | 0.045 |
| (150, 70) | 0.150 | 0.801 | 0.740 | 0.862 | 0.037 |
| (150, 70) | 0.250 | 0.801 | 0.724 | 0.874 | 0.045 |
| (250, 42) | 0.050 | 0.800 | 0.741 | 0.857 | 0.035 |
| (250, 42) | 0.150 | 0.800 | 0.742 | 0.858 | 0.035 |
| (250, 42) | 0.250 | 0.800 | 0.742 | 0.857 | 0.035 |

| *True Power = 90%* | | | | | |
|---|---|---|---|---|---|
| *Sample* | *Rho (ρ)* | *Mean* | *p(0.05)* | *p(0.95)* | *S.D.* |
| (60, 50) | 0.050 | 0.903 | 0.812 | 0.974 | 0.052 |
| (60, 50) | 0.150 | 0.903 | 0.811 | 0.974 | 0.052 |
| (60, 50) | 0.250 | 0.903 | 0.811 | 0.974 | 0.052 |
| (100, 30) | 0.050 | 0.902 | 0.832 | 0.958 | 0.039 |
| (100, 30) | 0.150 | 0.902 | 0.834 | 0.958 | 0.039 |
| (100, 30) | 0.250 | 0.902 | 0.834 | 0.957 | 0.039 |
| (150, 70) | 0.050 | 0.900 | 0.839 | 0.949 | 0.033 |
| (150, 70) | 0.150 | 0.901 | 0.850 | 0.943 | 0.028 |
| (150, 70) | 0.250 | 0.900 | 0.839 | 0.949 | 0.033 |
| (250, 42) | 0.050 | 0.900 | 0.854 | 0.939 | 0.026 |
| (250, 42) | 0.150 | 0.900 | 0.854 | 0.939 | 0.026 |
| (250, 42) | 0.250 | 0.900 | 0.854 | 0.939 | 0.026 |

NOTE: The Monte Carlo experiments that produced these results are described in Section IV and are identical to those underlying TABLE 2 with one exception: The clusters are now characterized by cross-sectional correlation. ICC continues to be represented by three values, $\rho = 0.050, 0.150,$ and $0.250$, but there is now a cross-sectional covariance term $= 0.030$, so that $\boldsymbol{\Omega}$ **is** described as below.

$$\boldsymbol{\Omega}_{NT\times NT} = \sigma^2 \times \begin{bmatrix} \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} & \cdots & 0.03\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \\ \vdots & \ddots & \vdots \\ 0.03\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} & \cdots & \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \end{bmatrix},$$