

# PAWS-CRD: User's Manual

---

## (Power analysis And Width of confidence interval for Sample size estimation for Cluster Randomized Design)

In a typical experiment, individuals are randomly assigned to one of two conditions, often an experimental condition and a control condition. In a cluster randomized design (CRD), groups of individuals are assigned to different conditions. For example, in classic educational effectiveness experiments, researchers compare the effectiveness of two teaching methods. The unit of randomization is the classroom. That is, instead of assigning individual children to one condition or the other, whole classrooms are assigned to be taught by one teaching method or by the other. The dependent variable is the academic performance of individual students in classrooms. In CRD terminology, the classroom is the cluster-level and student is the individual-level. The cluster and individual levels are sometimes referred to as the macro and micro level, respectively.

*A major challenge for researchers planning to use a cluster randomized design is that sample size estimation is more complex because there are two types of sample size: number of clusters (i.e., groups) and cluster size. Different combinations of these two types of sample sizes may provide the same chances to get a statistically significant result. Researchers may choose a sample size combination based on other desired characteristics, such as the least expensive combination. When the cost of adding a new cluster (e.g., new school) is much greater than the cost of finding a new individual (e.g., new student), it is preferred to reduce the sample by one school and recruit a few more students from each of the other schools to achieve the same power with a lower cost. However, if adding a new school to the study is inexpensive, researchers might prefer to gather data from many schools instead of recruiting new students from the schools already in the study.*

The purpose of the Power analysis And Width of confidence interval for Sample size estimation (PAWS-CRD) is to help researchers in the planning stage of their CRD studies. The program provides information about optimal sample size combinations based on several criteria: statistical power, accuracy in effect size estimation, and cost. Roughly speaking, statistical power in the context of a CRD study is the chance of getting a statistically significant result when there is a real mean difference between the experimental conditions. For example, when a new teaching method really increases students' academic achievement over the traditional method, the research study will draw a sample from the population and compare the new teaching method with the traditional method. The result of the research study may be significant or not, depending on how much sampling error is present. Statistical power is the probability getting a significant result, when the null hypothesis is false. Power is influenced by many factors but one that is directly under the researcher's control is sample size. Power is higher when sample size is higher. Therefore, the goal of power analysis in sample size estimation is to make sure that the sample size is large enough to have a high chance to get a significant result. PAWS-CRD estimates the cluster and individual-level sample size combination that results in the desired power with the lowest cost.

Traditionally, most researchers designing a study were primarily concerned with ensuring adequate power to detect the results they hope to find. Increasingly, however, researchers are focused on estimating effect sizes (ES) and the confidence intervals of their ES estimates (CI of ES). The effect size represents the

treatment effect. However, the effect size, such as Cohen's  $d$  (a.k.a., standardized mean difference), is a point estimate, which is subject to sampling error. To account for sampling error, researchers need to estimate a confidence interval of effect size with a specific degree of confidence, such as 95%. When the confidence interval is large, researchers are not sure whether replicated studies will obtain similar ES estimates. Thus, researchers would like to design their studies to make the CI of ES as narrow as possible. One way to reduce the width of CI of ES is to increase sample size. Thus, the program will estimate the cluster and individual-level sample size combination that provides the smallest width of CI of ES with the lowest cost.

One distinctive feature of the program is that it can inform researchers when it is desirable to have unequal proportions of treatment and control clusters. In other words, if the treatment condition is more expensive, the program will find the optimum proportion of treatment clusters by collecting more control clusters to minimize the budget while maintaining power or width of CI of ES. Furthermore, the program can find the sample sizes combination which provides the largest power or the smallest width of CI of ES when researchers have limited resources, such as a budget of \$10,000 for data collection.

Researchers with a limited budget are often forced to have small samples and thus low levels of power and a wide CI of ES. One way of increasing power or decreasing the width of the CI of ES is to introduce a covariate in the model. The covariate can be at either the cluster-level (e.g., school size or public/private school) or the individual-level (e.g., student's socioeconomic status). This program allows researchers to see how introducing a covariate in the model alters the optimal sample size combination to achieve a certain level power, accuracy in effect size estimation, or cost. However, one limitation of the program is that it assumes random assignment and thus the covariate is unrelated to treatment condition. That is, the means of the covariate variable are not different between treatment and control conditions. For example, socioeconomic status scores between old and new teaching method condition are equal. Another limitation, the program allows only one covariate in the model. However, as explained [elsewhere](#) in the manual, research designs with more than one covariate can be accommodated to some degree.

This program has a number of additional limitations. First, it is limited to two-level design. Second, in order to obtain very accurate results based on a priori Monte Carlo simulations, Mplus with the multilevel or combination add-on (Muthén & Muthén, 2007) must be installed on the same machine as PAWS-CRD. Without Mplus, researchers can still use this program to find the initial values, which will be reasonably accurate (and certainly better than random guesses) but not as accurate as the estimates from the Monte Carlo simulation. Researchers should refer to the [suggestion](#) in this manual to see the situations in which the initial values and the Monte Carlo results are not largely different. Third, the program is based on the ES using the individual-level standard deviation, not the total standard deviation, as is common in other programs (e.g., Optimal Design). I will describe the rationale for this decision when I explain the ES in CRD later.

**Before using the program, I highly recommend users to read the manual, or at least skim through it.** Misunderstandings about the use of the program may lead to erroneous estimations and problematic designs. To explain how to use the program, this manual covers some basic concepts, including basic CRD design, power, and width of CI of ES. Next, the basic program algorithms and their accuracy will be discussed. Then, an introduction to the use the program will be provided, along with five examples of

different research questions. Finally, recommendations for how to account for multiple covariates are given.

## Cluster Randomized Design

### Basic Concepts

As described above, CRD involves with nested data structure, such as students nested in schools. Students in the same school probably have similar experiences. This data structure is not appropriate for standard analysis, like independent  $t$ -test or one-way ANOVA, because these statistical techniques do not account for common experiences of individuals within clusters. To illustrate how CRD accounts for the common experience, I will partition the score in both ANOVA and CRD designs. In the ANOVA design, we can divide a score of each individual to three different parts,

$$Y_{ij} = \bar{Y}_{..} + \alpha_i + e_{ij} \quad (1)$$

In other words, the score of individual  $j$  in condition  $i$  ( $Y_{ij}$ ) can be divided to grand mean ( $\bar{Y}_{..}$ ), treatment  $i$  effect ( $\alpha_i$ ), and individual  $j$  error or his or her unique experience ( $e_{ij}$ ). We can see that the equation does not account for the cluster effect. In CRD design, a score of each individual can be divided into four different parts,

$$Y_{ijk} = \bar{Y}_{..} + \alpha_i + u_{ij} + e_{ijk} \quad (2)$$

In other words, the score of individual  $k$  in cluster  $j$  that has a condition  $i$  ( $Y_{ijk}$ ) can be divided to grand mean ( $\bar{Y}_{..}$ ), treatment  $i$  effect ( $\alpha_i$ ), cluster  $j$  error or common experience of individuals within cluster  $j$  ( $u_{ij}$ ), and individual  $k$  error or his or her unique experience ( $e_{ijk}$ ).

The error in ANOVA design varies across individuals. Thus, we can find the amount of variation of the error term, which is error variance or mean on squared error. However, when comparing to ANOVA design, CRD divided the error in ANOVA design to two parts: cluster error and individual error. Cluster error and individual error have their own variation, which are cluster error variance ( $\tau$ ) and individual error variance ( $\sigma^2$ ). Cluster error variance and individual error variance are not necessarily the same or even of the same magnitude. For example, it is possible that in a particular dataset the differences in student socioeconomic status is substantial across schools, but the differences among students within each school is trivial. In other words, cluster error variance in socioeconomic status can be greater than individual error variance. Of course, it is possible for individual error variance to be much larger than cluster error variance. The sum of both error variances is equal to the total error variance in the ANOVA design. Thus, we can find how much the error variance allocate to cluster error variance that is known as intraclass correlation ( $\rho$ ),

$$\rho = \frac{\tau}{\tau + \sigma^2} \quad (3)$$

In the design that researchers would like to compare two condition means, the score equation can be written as

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + u_j + e_{ij} \quad (4)$$

In other words, the score of individual  $i$  from cluster  $j$  ( $Y_{ij}$ ) is the sum of the intercept ( $\gamma_0$ ), the treatment condition effect ( $\gamma_1$ ), the group error ( $u_j$ ) associated with membership in cluster  $j$ , and the individual error

( $e_{ij}$ ) associated with individual  $i$  in cluster  $j$ . This equation is similar to analyze independent  $t$ -test or ANOVA with two means comparison, which shown in the Equation 1, by regression analysis. The intercept and treatment condition effect are interpreted differently, depending on how the treatment variable is coded. For usefulness in defining ES, I develop the two-group CRD using dummy coding for the treatment variable: 1 for the treatment condition and 0 for the control condition. The treatment condition effect ( $\gamma_1$ ) is the difference between the treatment condition and control condition means. The intercept ( $\gamma_0$ ) is the control condition clusters mean. The variance of the difference between treatment and control condition means estimate is

$$\text{Var}(\hat{\gamma}_1) = \frac{\sigma_Y^2 + n\tau_Y}{nkp(1-p)} \quad (5)$$

where  $k$  is number of clusters,  $n$  is the cluster size,  $p$  is the proportion of treatment condition clusters,  $\tau_Y$  is the cluster error variance or the variance of cluster-level error term ( $u_j$ ) and  $\sigma_Y^2$  is the variance of individual-level error term ( $e_{ij}$ ).

In ANOVA, we can introduce a covariate in the model to decrease error variance and then increase statistical power. In the two-group CRD model, we can introduce a covariate and hope that error variances are reduced. However, cluster and individual error variance reduce in different degrees, based on how much a covariate explains error variances at each level: between-cluster and within-cluster effects. For example, socioeconomic status can explain academic achievement in both the cluster and the individual level. The within-cluster effect is the degree to which socioeconomic status explain the variability of academic achievement within schools. However, the between-cluster effect is the degree to which socioeconomic status differences across schools explains the variability of academic achievement differences across schools. Therefore, a covariate score can be divided into two portions: difference between- and within-group,

$$Z_{ij} = \bar{Z}_j + (Z_{ij} - \bar{Z}_j) \quad (6)$$

where  $Z_{ij}$  is the covariate score of individual  $i$  from cluster  $j$  and  $\bar{Z}_j$  is the average covariate score across individuals in cluster  $j$ . Including a covariate effect in the Equation 4, each dependent variable score can be divided as,

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + \gamma_B \bar{Z}_j + \gamma_W (Z_{ij} - \bar{Z}_j) + u_j + e_{ij} \quad (7)$$

where  $\gamma_B$  is the group-level effect toward the dependent variable on the covariate,  $\gamma_W$  is the individual-level effect toward the dependent variable on the covariate,  $u_j$  is the cluster-level error that accounts for the part of the score that cannot be explained by treatment variable and cluster-level covariate, and  $e_{ij}$  is the individual-level error that cannot be explained by the individual-level covariate. Other terms are defined in the Equation 4 and 6. This method of putting a covariate in the equation is also known as group-mean centering (Enders & Tofighi, 2007). Given the full range of options afforded by multilevel modeling, the individual-level covariate regression coefficient ( $\gamma_W$ ) does not have to be equal across clusters. However, this program will focus only on a covariate that has constant effects across groups.

Also, both error terms are varied but to a lesser degree compared to the model without a covariate. Let  $\tau_{(Y|Z)}$  and  $\sigma_{(Y|Z)}^2$  be the cluster- and individual-level error variance after partialing out the effect of

covariate  $Z$ . Then, the proportions of error variances explained by covariate  $Z$  (Byrk & Raudenbush, 2002) are

$$R_B^2 = \frac{\tau_Y - \tau_{(Y|Z)}}{\tau_Y} \quad (8)$$

$$R_W^2 = \frac{\sigma_Y^2 - \sigma_{(Y|Z)}^2}{\sigma_Y^2} \quad (9)$$

where  $R_B^2$  is the proportion of variance explained at the cluster level and  $R_W^2$  is the proportion of variance explained at the individual level. As mentioned above,  $R_B^2$  and  $R_W^2$  are not necessarily equal.

In addition, the overall variance of a covariate can be divided into two levels: cluster-level ( $\tau_Z$ ) and individual-level ( $\sigma_Z^2$ ). The intraclass correlation of the covariate ( $\rho_Z$ ) is

$$\rho_Z = \frac{\tau_Z}{\tau_Z + \sigma_Z^2} \quad (10)$$

The variance of the difference between treatment and control condition means estimate is

$$\text{Var}(\hat{Y}_1) = \frac{\sigma_{(Y|Z)}^2 + n\tau_{(Y|Z)}}{nkp(1-p)} \quad (11)$$

However, a covariate cannot only be an individual property but also a property of clusters, such as teaching performance in a classroom-students data structure. In this case, a covariate will have only cluster-level effect and only explain variance at the cluster level.

### Effect Size and its Confidence Interval

In independent  $t$ -test, the most popular effect size index is Cohen's  $d$  or standardized means difference, which is computed by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \quad (12)$$

where  $\bar{X}_1$  and  $\bar{X}_2$  are group means of condition 1 and 2, respectively, and  $s_p$  is pooled standard deviation from both groups.

However, Cohen's  $d$  cannot be used directly in the CRD context because there are many kinds of standard deviations: overall, cluster-level, and individual-level standard deviations (Hedges, 2007). As the benefit of ES is to compare with other studies, these standard deviations are appropriate in different situations. When the lower level of CRD is the individual level and other studies are usually a single-site study, the individual-level standard deviation is appropriate. When the higher level of CRD is individual level, such as comparing individuals who have one or many GRE scores, the group-level standard deviation is appropriate. However, when most of other studies in the area are large surveys that report the overall standard deviation and ignore the fact that people are nested in natural groups, the overall standard deviation is appropriate. As CRD most often used when the lower level is the individual level, I will focus on only Cohen's  $d$ , which uses the individual-level ES. These individual-level ES ( $d_W$ ) can be written as

$$d_w = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_Y} = \frac{\gamma_1}{\sigma_Y} \quad (13)$$

where  $\gamma_1$  is the regression coefficient when the grouping variable was coded as dummy variable (0 or 1) in the Equation 4 and 7 and where  $\sigma_Y$  is the individual-level standard deviation.

There are at least two options available for computing CI of ES using individual-level standard deviation. First, we can use the formula provided by Hedges (2007). However, this method is not easy to generalize when a covariate is introduced in the model. Another method of computing CI of ES takes advantage of features found in many structural equation modeling packages (Cheung, 2009). This method can analyze the CI of any statistic that can be expressed by formulas. In most packages, a latent variable, called the phantom variable, is specified to have zero variance. An arbitrarily chosen observed variable is specified to regress onto the phantom variable. Next, the regression coefficient of the phantom variable represents the statistic needed by linear or nonlinear constraints from other parameters in the model. After analysis, the standard error of the regression coefficient is obtained and can be used to construct the CI by normal approximation. Therefore, to construct CI of ES in CRD, the regression coefficient can be constrained as the regression coefficient from treatment variable to dependent variable divided by the appropriate standard deviation of the dependent variable. This program will use Mplus (Muthen & Muthen, 2007) to obtain the CI of ES in CRD.

## Power and Width of CI of ES

### Basic Concepts

Before going through the details of the program, let's talk about the criteria used for sample sizes estimation: power analysis and accuracy in parameter estimation. The power analysis is the most popular approach that finds a chance of finding statistically significant results given the real effect in the population. The power is positively related to effect size. The larger the effect size, the greater the power of analysis. Also, the more sample sizes provides the more power of analysis. Usually, researchers would like an analysis to have enough power (e.g., .80) to test their research hypotheses. Therefore, given the effect size, the researchers would like to collect enough number of people for a specified power. If researchers do not have enough power in their samples and get a result that is not statistically significant, the researchers cannot conclude that the treatment is not effective or the sample sizes are not large enough to detect the real effect. The planning in CRD is much more crucial because the power in CRD is usually less than single-level design. Also, the increases in cluster size and number of clusters have different effects. A large number of total individuals do not mean that the design has high enough power. The large number people in each cluster but the lower number of clusters will give a very small power. The program will pick an efficient combination of sample sizes which provides a specified power. The effect size for finding a power is the Cohen's  $d$  that is standardized by individual standard deviation, as shown above. The further details of power analysis can be found in Cohen (1988) or the 4-page introduction of power analysis by Cohen (1992).

Another approach is to find the sample sizes that have enough accuracy in parameter estimation. This approach is mostly found in the survey research. For example, a political scientist would like to estimate the proportion of US people choosing a president candidate. He or she would like to make the estimation accurate enough in the population inference (e.g.  $\pm 2\%$ ) by collecting enough sample sizes (e.g. 1000 people). The further details of finding sample size for inferring a descriptive statistic can be found in the classic book of Cochran (1977).

However, most educational and social variables have arbitrary scale that researchers cannot infer the magnitude of difference from a difference score. Therefore, the parameter estimation of the raw score means difference is not the researchers' interests. Recently, Wilkinson and the Task Force of Statistical Inference (1999) encouraged researchers to use an effect size index and its interval estimation of the effect size. The confidence interval (CI) of effect size (ES) provides several advantages over traditional significance testing (see Thompson, 2002). It does provide all information the significance testing have as well as the additional information, such as effect size, accuracy of the effect size estimation, or potential application in accepting the null hypothesis. As most statistics students have been warned repeatedly by their instructors, significance testing cannot be used for that purpose.

As a consequence, a researcher may want to find the sample sizes to ensure that the effect size is accurate enough (i.e., the width of CI of ES is narrow enough). For example, government would like to ensure whether a new teaching method is effective. The sample ES may show that the teaching method is effective, such as the result is highly significant and the effect size is quite large. However, how do we know that this effective sample ES is robust in the replication study or the real world implementation? Thus, the sample ES should be accurate enough for inference.



One of the complexities in finding width of CI of ES is that width of CI of ES is subject to sampling error. The obtained width in the real data will deviate (positively or negatively) from the specified width. Therefore, the obtained sample size does not guarantee that the obtained width of CI is equal to the desired width. It might be said that the probability of getting samples which the width of CI is less than or equal to the desired width is approximately 50%. Thus, researchers might want to increase their sample size somewhat so that the probability of getting a larger width of CI than desired is reduced. The probability that obtained samples have a CI of ES width less than or equal to desired width is known as degree of certainty. Researchers may want to estimate sample sizes which have 80% confidence (i.e., degree of certainty) that the CI of ES is not larger than a specified width (e.g., 0.5). This program will allow researchers to estimate the sample sizes based on the width of CI of ES that is standardized by individual-level standard deviation only. The further details can be found on Kelley and Rausch (2006). The Maxwell, Kelley, and Rausch (2008) article is a good review in comparing both power and accuracy in parameter estimation approach for finding sample sizes.

### Power and Width of CI of ES in CRD

There are at least two approaches to find power and width of CI of ES. First, the specified power or width of CI of ES can be transformed to the variance of the treatment effect by

$$\text{Optimal Var}(d_w) = \left( \frac{d_w}{z_{1-\alpha/2} - z_{1-\text{power}}} \right)^2 \quad (14)$$

for a given power and

$$\text{Optimal Var}(d_w) = \left( \frac{\omega}{2z_{1-\alpha/2}} \right)^2 \quad (15)$$

for a specified width of CI of ES, where  $d_w$  is ES estimate using individual-level standard deviation,  $z_x$  represents quantiles in the normal distribution, and  $\omega$  is width of CI of ES using individual-level standard deviation.

Assuming that the individual-level standard deviation is 1, the treatment effect and the individual-level effect size is equal. Therefore, we can solve for a combination of number of clusters, cluster size, and proportion of treatment clusters using the Equation 5 or 11, by letting the variance equal to the optimal variance given from the specified power or width of CI of ES.

The program will find the combination which provides the lowest budget. The lowest budget can be defined by

$$C = C_T + C_C \quad (16)$$

where total treatment ( $C_T$ ) and total control cost ( $C_C$ ) are

$$C_T = k_T \times (\text{treatment cluster cost} + (n \times \text{treatment individual cost})) \quad (17)$$

$$C_C = k_C \times (\text{control cluster cost} + (n \times \text{control individual cost})) \quad (18)$$

If researchers would like to specify some properties in advance (e.g., cluster size), the program just substitute the value in the equation in advance and find the other unknowns.

This approach is just the approximation of the sample sizes because it makes two assumptions: individual-level standard deviation equal to 1 and the effect size varied as normal distribution. The more accurate approach is to make the sampling distribution of the effect size instead of the raw score mean differences. The sampling distribution of effect size is noncentral  $t$ . When the degree of freedom is low and the effect size is large, the normal distribution is not the good approximation of noncentral  $t$  distribution. Therefore, the program will use the first approach for just finding the initial estimation of sample sizes. The second approach will adjust the sample sizes to the more accurate one.

As the second approach, a priori Monte Carlo simulation will be used to estimate power or CI of ES. A priori Monte Carlo simulations build large numbers of samples from specified parameters. Next, for power analysis, the samples are used to calculate the desired statistics and check how many samples are statistically significantly different from the null hypothesis value. For the accuracy in ES estimation, the width of CI of ES is analyzed from each sample and the average width or the degree of certainty (i.e., the percentage of samples that have the width of CI of ES less than the specified width) will be analyzed from all simulated samples. The program will check whether the initial sample sizes give the accurate power or width of CI of ES and give the lowest cost compared to other samples. In this process, a large number of samples are required. Muthén and Muthén (2002) used 10,000 hypothetical samples for the analyses. In my opinion, however, the sample size estimation is not necessary to be very accurate. Using 10,000 hypothetical samples will consume analytic time too much, especially generating samples with multiple levels. I suggest that 1,000 hypothetical samples are enough for sample size estimation.

For the limited budget approach, the program will find various combinations of sample sizes by solving the Equation 16 for a given budget. Next, the sample sizes are substituted in the Equation 5 or 11. The program will find the sample sizes combination which provides the lowest variance of treatment effect (i.e. largest power and lowest width of CI of ES).

## Program Accuracy

I used PAWS-CRD to replicate the results obtained from PINT 2.2 that accompanies the Snijders and Bosker (1993) article. PINT is preferred because it can be used to estimate standard errors in most two-level models. However, it requires many steps to obtain the standard error for CRD, such as calculating the variance of the treatment variable, the variance of the covariate in both levels, the mean of the treatment variable, and the error variance of the dependent variable in both levels.

I validated PAWS-CRD based on 300 situations from combinations of five variables.

1) I used five *methods to find the sample sizes*:

- (a) to achieve power of .80 and minimize budgets,
- (b) to achieve the width of 0.2 and minimize budgets,
- (c) to achieve the width of 0.5 and minimize budgets,
- (d) to maximize power given \$500 budget, and
- (e) to maximize power given \$1,000 budget.

I used a power of 0.8 based on Cohen's (1988) guideline. I assumed that researchers would not like the width of CI greater than 0.5. I also assumed that, for pragmatic reasons, researchers would not like to have a CI of ES that is too narrow. Thus, I considered only CI of ES's with widths of 0.2 and 0.5. In the simulations, the budget was arbitrarily set to \$500 or \$1000 so that when the individual cost is \$1, the budget would be sufficient for 500 or 1000 individuals.

2) I specified the *intraclass correlation of the dependent variable* as either 0.05 or 0.25. These numbers are inspired by the findings in Hedges and Hedberg (2007) that academic achievement had the intraclass correlation of 0.25 within classrooms and that many other psychological constructs have an intraclass correlation of 0.05 within classrooms.

3) I specified the *effect size of the treatment variable* on the dependent variable to be either 0.2 or 0.5, which corresponds to Cohen's (1988) labels of small and medium effect sizes, respectively. Note that all effect sizes are based on individual-level standard deviations.

4) I specified three *group costs*: None, \$5, and \$10. The choices of \$5 and \$10 are to specify that the group costs are five and ten times that of the individual cost, which specified as \$1.

5) I used five *covariate characteristics*:

- (a) no covariate,
- (b) an individual-level covariate explaining 13 percent of individual error variance,
- (c) a covariate with intraclass correlation of 0.5 explaining 13 percent of both cluster and individual error variance,
- (d) a covariate with intraclass correlation of 0.25 explaining 13 percent of both cluster and individual error variance, and
- (e) a cluster-level covariate explaining 13 percent of cluster error variance.

The value of 13 percent was chosen to correspond to a medium effect size of  $f^2 = .15$  in a multiple regression model (Cohen, 1992).

Given the five variables described above, there are  $5 \times 2 \times 2 \times 3 \times 5 = 300$  combinations that were evaluated. I validated PAWS-CRD by finding the sample sizes given each situation. Next, I used PINT to estimate the standard error of the treatment effect. The standard error will be used to calculate the power and width of CI of ES using the Equation 14 and 15, respectively. I also used PAWS-CRD to find the power, width of 95% CI of ES, and width of 99% CI of ES, based on the normal approximation method, which is used to calculate starting value, and the a priori Monte Carlo simulation method.

The power, width of 95% CI and width of 99% CI provided PINT and the normal approximation method are different from each other by no more than 0.001 across 300 situations, which is comparable to rounding error. Thus, the method provided by Snijders and Bosker (1993) and the normal approximation method are essentially the same. PAWS-CRD calculates the starting values accurately. The difference between the normal approximation, including the Snijders and Bosker (1993) method, is discussed in next section.

### **Difference in Sample Size Estimation between Two Approaches**

A priori Monte Carlo Simulation will take a long time to create and analyze a thousand samples. While I tested the program, I also looked for the discrepancy between the normal approximation method and the a priori Monte Carlo simulation method. As described above, the benefit of the a priori Monte Carlo simulation method is that 1) it accounts for sampling error of intraclass correlation in the dependent variable and 2) it treats the covariate as a random effect.

The power, width of 95% CI and width of 99% CI for the normal approximation method and the a priori Monte Carlo simulation method are mostly similar. For power, the differences between two methods range from -.07 to .04. The positive sign indicates that the approximate method is greater. For 95% and 99% CI of ES, the differences range from -2.07 to 0.04 and -2.72 to 0.06, respectively. Thus, in some situations, the approximate method underestimates the width of the CI of ES by a large amount. The discrepancies are mostly affected by the type of covariate, which is shown in Table 1. The covariate with an intraclass correlation of 0.05 is the only condition that resulted in large discrepancies. Upon further exploration, it appears that an intraclass correlation of 0.05 only resulted in large discrepancies when the dependent variable had an intraclass correlation of 0.25 and the total sample size was less than 500. However, more systematic explorations are needed to pinpoint exactly which situations produce large discrepancies between the two methods.

However, according to this study, some recommendations can be made. To find power, the starting values can provide quite accurate results. For the width of CI of ES, generally, researchers can use the program to estimate sample size combinations. However, when PAWS-CRD recommends low total sample sizes (i.e., total sample size < 500) and the intraclass correlation of the covariate is less than the intraclass correlation of the dependent variable in some degree (e.g., .05), the a priori Monte Carlo simulation is highly recommended. For researchers who need a high degree of accuracy, the a priori Monte Carlo simulation is recommended in all cases. I admit that, in some situations, the a priori Monte Carlo simulations needed 12 or more hours to be completed. I am still working on finding ways to speed up the algorithm.

Table 1

*Difference between the approximate method and a priori Monte Carlo method in estimating power, 95% CI of ES, and 99% CI of ES across five conditions.*

Type of Covariate	Difference in Power				Difference in 95% CI of ES				Difference in 99% CI of ES			
	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max
No Covariate	-0.005	0.015	-0.073	0.014	0.001	0.007	-0.009	0.028	0.002	0.009	-0.012	0.038
Individual-level Covariate	-0.009	0.015	-0.044	0.012	-0.004	0.006	-0.024	0.023	-0.006	0.008	-0.032	0.030
Covariate with ICC of 0.05	-0.004	0.014	-0.044	0.035	-0.128	0.319	-2.069	0.003	-0.168	0.419	-2.719	0.004
Covariate with ICC of 0.25	-0.008	0.015	-0.049	0.021	-0.007	0.006	-0.026	0.008	-0.009	0.007	-0.034	0.011
Group-level Covariate	-0.009	0.014	-0.063	0.006	0.002	0.009	-0.027	0.044	0.003	0.012	-0.036	0.057
Total	-0.007	0.015	-0.073	0.035	-0.027	0.150	-2.069	0.044	-0.036	0.198	-2.719	0.057

*Note.* Positive sign indicates that the approximate method is greater. Each type of covariate contains 60 situations. CI = Confidence Interval. ES = Effect Size. ICC = Intraclass Correlation.

## Program Illustration

The program is limited to a simple case of CRD, which is

- 1) The model involves only two condition comparison, such as experiment and control conditions.
- 2) CRD involves only two levels.
- 3) As the definition of CRD, treatment variable are in the cluster level (i.e., macro level).
- 4) As shown above, treatment effect size is defined by the individual-level standard deviation only.
- 5) Only one covariate can be introduced in the model. However, the method to aggregate multiple covariates is proposed later.
- 6) Covariates are not different across conditions, which is satisfied when clusters are randomly assigned to each condition.

The program can be divided to two different parts: post hoc and a priori. Post hoc analyses are used to find the power or width of CI of ES when the ES of the treatment effect using the individual-level standard deviation and sample sizes are known. In contrast, a priori analyses are used to estimate sample sizes given a specific level of power (or width of CI of ES) and ES of treatment effect using individual-level standard deviations. Users can choose either post hoc or a priori by choosing the tab below the menu bar.

## Post Hoc Analysis

Researchers may want to know the power or CI of ES based on the results they obtained after the experiment was conducted. Some researchers may use post hoc power to understand their analyses further, whether their results were statistically significant or not. In general, the post hoc power of a nonsignificant result will be low and the post hoc power of a significant result will be high. Post hoc power has a monotonic relationship with the  $p$  value (Hoenig & Heisey, 2001). Thus, I do not recommend researchers to use post hoc power as a supplemental analysis. For computing CI of ES, I will recommend researchers to use the [method discussed above](#). My aim in building the post hoc part is to allow researchers to explore the data and use the result to help in the design future research, as Howell (2007) recommended as a good way to use post hoc power. In this part, two examples are shown to illustrate the model with and without a covariate.

### Example 1

Let's say a researcher studies the effect of group-based intervention on depression level. This example is modified from King et al. (2002) study. Eighty-four therapists were randomly assigned to be trained to administer cognitive behavioral therapy or not. Each therapist treated four patients. Each patient completed the Beck Depression Inventory. Within therapist groups, there is an intraclass correlation .11 (adjusted for the intervention effect, the ICC becomes 0.013). Patients with therapists trained in cognitive behavioral therapy have an average score of 17.5 (SD = 9.6) on the Beck Depression Inventory. In the control condition, the average score was 16.6 (SD = 11.5).

### Inputs

To estimate the post hoc analysis, the **Post Hoc** tab is selected, as shown in the red box of Figure 1. Next, clear the **Activate Covariate** checkbox in the list of checkboxes as shown in the blue box of Figure 2. We can see that the top left group box, the **Input** groupbox, is the only activated group box.

Figure 1. Checkbox characteristics for the Example 1

In the model without a covariate, five input boxes, as shown in the green box in Figure 1, need to be filled as shown in the Figure 2. The **Number of Clusters in Treatment Condition** is the number of therapists who were trained in cognitive behavioral therapy. In this example there are 42. The **Number of Clusters in Control Condition** is the number of therapists who were not trained, which is also 42. The **Number of Participants in Each Cluster** or the cluster size is 4, which is the number of patients per therapist.

The **Intraclass Correlation of Y** in this program is the proportion of cluster-level error variance to total error variance. Thus, the treatment variance is not included, so the number of 0.013 is put in the program.

**Effect size (Using Individual-Level SD)** is required some calculations. First, the standard deviations from both conditions are pooled by

$$\text{Pooled Variance} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{(168 - 1)(9.6)^2 + (168 - 1)(11.5)^2}{(168 - 1) + (168 - 1)} = 112.21$$

Second, find the individual-level variance. Because the pooled variance includes the variance from treatment effect, the intraclass correlation unadjusted for the treatment variable is used.

$$\rho_{\text{Unadjusted}} = \frac{\text{Treatment Variance} + \tau}{\text{Treatment Variance} + \tau + \sigma^2}$$

$$1 - \rho_{\text{Unadjusted}} = \frac{\sigma^2}{\text{Treatment Variance} + \tau + \sigma^2}$$

The denominator of the right expression is the total variance, so

$$1 - 0.11 = \frac{\sigma^2}{112.21}$$

After solving the equation, the individual-level variance is 99.86 and individual-level standard deviation, the square root of the variance, is 9.99. The effect size is the mean difference between two conditions divided by the individual-level standard deviation,  $(17.5 - 16.6)/9.9 = 0.09$ . Thus, 0.09 is entered in the **Effect Size (Using Individual-Level SD)** box.

After that, researchers may choose either run a Monte Carlo simulation or not by checking or clearing the **Use A Priori Monte Carlo Simulation** checkbox in the top right of the checkbox list, as shown in the blue box of the Figure 1. When running a priori Monte Carlo Simulation, users may change the number of replicated samples used in Edit → Options. I recommend using 1,000 replicated samples to obtain an accurate result. The number of replication can be changed by going to **Edit → Options** in the menu bar. After everything is set up, click “Calculate” button.

### Outputs

As the first example, I will run the estimation with and without a priori Monte Carlo simulation. When running the model with a priori Monte Carlo simulation, the result also shows in the top right group box, as shown in Figure 2. The power is 0.150. The width of 95% and 99% CI of ES is 0.441 (range from -0.130 to 0.310) and 0.579 (range from -0.200 to 0.038), respectively.

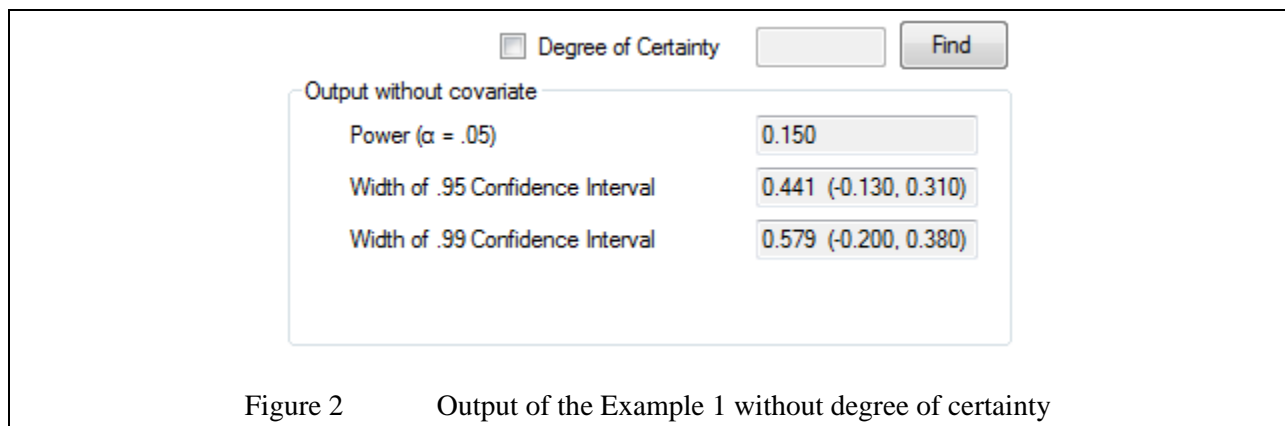
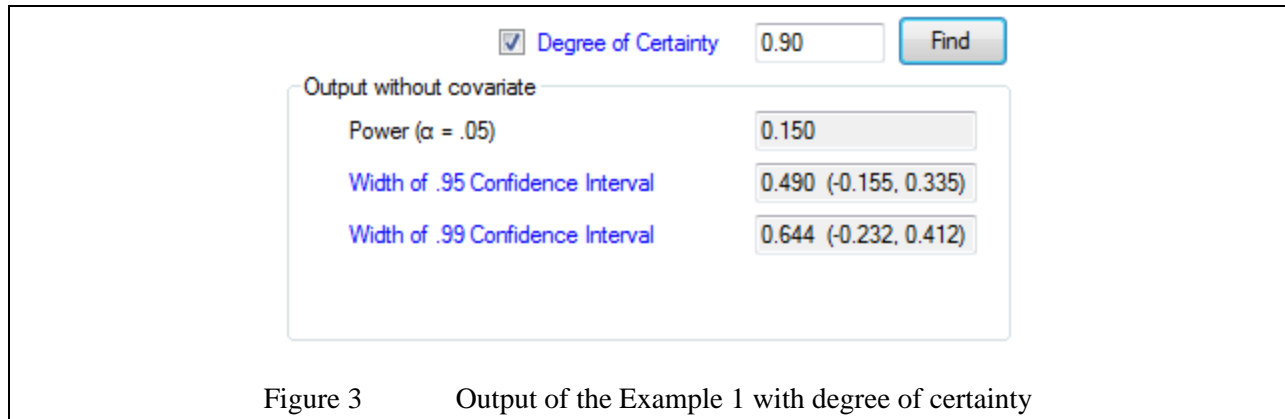


Figure 2 Output of the Example 1 without degree of certainty

The degree of certainty function is available by activating **Degree of Certainty** checkbox. After that, the number of degree of certainty can be specified in the right box of the text. Next, click **Find**. As shown in Figure 3, I specified the degree of certainty as 0.90 and click “Find.” The text colors of the widths of 95% and 99% Confidence Intervals are changed to blue. The boxes on the right of the width of CI provide the width of 95% and 99% CI given 90% degree of certainty. In this context, 90% of similar studies will have the width of 95% CI of ES at most 0.490 (corresponding to the range from -0.155 to 0.335) and the width of 99% CI of ES at most of 0.644 (corresponding to the range from -0.232 to 0.412).





When running the model without a priori Monte Carlo simulation, the result shows in the same box as in the Figure 2, but the numbers are changed slightly. The power is 0.124. The width of 95% and 99% CI of ES is 0.439 (range from -0.129 to 0.309) and 0.577 (range from -0.198 to 0.378), respectively. The degree of certainty function is not available. For all analyses, users may click [File → Show Summary](#) to show the output in text file.

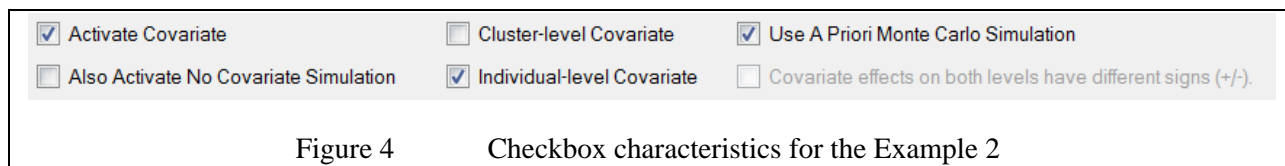
## Example 2

From the previous example, the researcher would like to use pretest of the depression scale as a covariate. She expects that the covariate is not varied across groups because the groups are formed after the pretest phase. Thus, the pretest of the depression level is individual-level covariate. She expects that the pretest accounts for at least 50% of variance of individual error variance of the posttest depression scores.

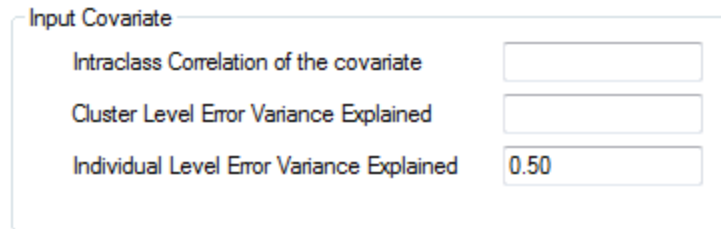
### Inputs

As in the Example 1, the **Post Hoc** tab is used. In the **Input** group box, the **Number of Clusters in Treatment Condition** is 42. The **Number of Clusters in Control Condition** is 42. The **Number of Participants in Each Cluster** is 4. **Intraclass Correlation of Y** is 0.013. The **Effect Size (Using Individual-Level SD)** is 0.09.

As shown in Figure 4, the covariate can be introduced in the program by checking **Activate Covariate** in the list of checkboxes. After that, the **Individual-level covariate** checkbox is available. Check on **Individual-level covariate**. Notice that the **Input Covariate** group box, as shown in Figure 5, the **Intraclass Correlation of the Covariate** and **Cluster Level Error Variance Explained** are inactivated.



Then, put 0.50 in the **Individual Level Error Variance Explained** box to indicate that 50% of the error variance is explained by the pretest. Again, users can check or clear **Use a Priori Monte Carlo Simulation** to activate or inactivate a priori Monte Carlo simulation, respectively.

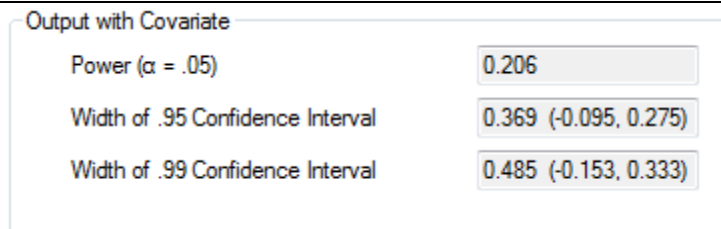


Input Covariate	
Intraclass Correlation of the covariate	<input type="text"/>
Cluster Level Error Variance Explained	<input type="text"/>
Individual Level Error Variance Explained	0.50

Figure 5 Covariate input for the Example 2

### Outputs

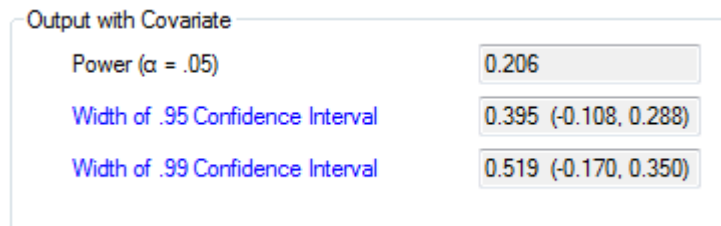
I will show only the result of the a priori Monte Carlo simulation method. As shown in Figure 6, the obtained power is 0.206. As shown in the Example 1, the power is 0.150 without the covariate. Thus, including the covariate increases the power only 0.05. The widths of 95% and 99% CI of ES are 0.369 (range from -0.095 to 0.275) and 0.485 (range from -0.153 to 0.333), respectively. Compare to the Example 1, the reductions of the widths of CI of ES are small.



Output with Covariate	
Power ( $\alpha = .05$ )	0.206
Width of .95 Confidence Interval	0.369 (-0.095, 0.275)
Width of .99 Confidence Interval	0.485 (-0.153, 0.333)

Figure 6 Output of the Example 2 without degree of certainty

The degree of certainty function is available by checking **Degree of Certainty**. After checking, the value of degree of certainty can be specified in the right box of the text. Next, click **Find**. For illustration, I specified the degree of certainty as 0.80. As shown in Figure 7, the text color of the widths of 95% and 99% CI is changed to blue. The boxes on the right of the width of CI provide the width of 95% and 99% CI given 80% degree of certainty. In this context, 80% of similar studies will have the width of 95% CI of ES at most 0.395 (corresponding to the range from -0.108 to 0.288) and the width of 99% CI of ES at most of 0.519 (corresponding to the range from -0.170 to 0.350).



Output with Covariate	
Power ( $\alpha = .05$ )	0.206
Width of .95 Confidence Interval	0.395 (-0.108, 0.288)
Width of .99 Confidence Interval	0.519 (-0.170, 0.350)

Figure 7 Output of the Example 2 with degree of certainty

## A Priori Analysis

The aim of a priori analysis is to find sample size combination based on different approaches. PAWS-CRD can compute sample size combination based on three criteria.

- 1) Researchers specify the power, the width of CI of ES, or the degree of certainty in advance and try to find the sample size combination with the lowest total sample sizes.
- 2) Researchers specify the power, the width of CI of ES, or the degree of certainty in advance and try to find the sample size combination with the lowest total cost.
- 3) Researchers have a limited budget and would like to find a sample size combination that provides the highest power, which also provides lowest width of CI of ES.

To specify the design characteristics, researchers may use guideline of Cohen's  $d$  (Cohen, 1988) or intraclass correlation (Hedges & Hedberg, 2007). I do recommend that researchers should find the design characteristics based on the previous studies as possible, which provides the better estimation. In the second and third criteria, researchers need to specify treatment group cost, control group cost, treatment individual cost, and control individual cost. The definitions of the costs are [provided above](#). The cost is not limited to data collection cost. The cost includes everything in the research design, such as an intervention cost, a cost to hire more research assistants, a cost to collect feedback from participants (even though it is not related to research purposes), etc.

When researchers specify treatment costs different from control costs, the program probably provides number of treatment and control groups differently. When the numbers of treatment and control groups are different, the result of the statistics may be biased as parametric assumptions are violated, such as normality or homogeneity of error variances. Thus, when researchers concern about the violation of parametric assumptions, researchers may specify the proportion of treatment groups as 0.50 (i.e., the number of treatment and control groups are equal) in advance.

The number of individuals and the number of groups may be specified in advance also. For example, researchers need to collect all students from each classroom. Thus, the number of individuals in each group cannot be changed and need to be specified as 25. Note that, when the number of groups is specified as a very low number, the program may be not able to find a sample size combination that provides a level of power equal to the specified level. Although the limit of power when increasing the number of groups approaches 1.0, the limit of power when increasing the group size is not 1.0.

The a priori Monte Carlo simulation in an a priori analysis takes a long time. To find a sample size combination for a specified degree of certainty, I highly recommend that researchers should find the starting values of sample size combination for a given width of CI of ES first. Then, use the number of individuals from the starting values as a guess for the number of individuals in the optimal sample size. Next, specify the estimated sample size in the design characteristics and use a priori Monte Carlo simulation to find the optimal sample size combinations for the specified degree of certainty based on number of clusters and proportion of treatment clusters. When the number of individuals in each cluster is specified in advance, the time that the a priori Monte Carlo simulation takes will decrease for at least three times!!! Three examples are illustrated to show 1) the model given a power using the second criterion, 2) the model given a degree of certainty using the second criterion, and 3) the model using the third criterion.

### Example 3

Educational researchers would like to implement a new teaching method comparing with the traditional method. Classrooms are randomly assigned to the new teaching method and traditional method. The researchers would like to achieve power of 0.8. They think that the meaningful effect size for this study is 0.2, which is considered as low effect size based on Cohen's (1988) guideline. A good estimate of the within classroom intraclass correlation of the achievement variable is 0.25 based on the Hedges and Hedberg's (2007) guideline. The cost to collect the data from each student is \$2 for both teaching methods conditions. The new teaching method provides many more activities than the traditional method. Therefore, the expenditures for implementing the new teaching method and the traditional method to each classroom are \$600 and \$300, respectively. The classroom size is 25 students.

### Inputs

First, click the **A Priori** tab on the top left of the window. Because this design does not have a covariate, make sure that the **Activate Covariate** at the checkbox list is clear. Next, in order to specify a criterion to estimate a sample size combination, click the **Define** button below the **Find Sample Size Criterion** group box. The new pop-up window will appear as Figure 8.

Because the researchers would like to find the least expensive sample size combination for a given level of power, the second criterion, **Minimize the Cost with a Specified Power or Width**, in the **Input Criterion** group box will be chosen. In the **Pre-Defined Sample Size** group box, researchers will consider about pre-specified characteristics of the result of sample size combination. In this example, researchers specify the number of individuals in each group equals 25 in advance. Thus, check **Number of Individuals in Each Cluster** and put 25 in the right box. Next, **Cost Information** is specified. The treatment group cost is 600. The control group cost is 300. The individual costs for both treatment and control conditions are 2.

In the **Desired Power or Width of Confidence Interval of Effect Size** group box, researchers will specify the desired sample size characteristics based on either power or width of CI of ES. In the **Width of CI of Effect Size** option, researchers can decide based on either the average width of CI of ES or the degree of certainty. In this example, the **Power** will be chosen and specified as 0.80. After everything is set, click **OK**.

The result of sample size characteristics specification will be shown in the **Finding a Sample Size Criterion** textbox, as shown in the Figure 9. Next, the **Effect Size (Using Individual-Level SD)** is set as 0.2 and the **Intraclass Correlation of Y** is set as 0.25. Make sure that the **Use a priori Monte Carlo Simulation** is checked. Finally, click **Calculate** to run an analysis.

**Criteria for Finding Sample Size**

**Input Criterion**

- ☐ Minimize the Total Number of Individuals with a Specified Power or Width
- ☒ Minimize the Cost with a Specified Power or Width
- ☐ Maximize the Power or Minimize the Width with a Limited Budget

**Pre-Defined Sample Size**

Do you wish to specify the sample size results in advance?

- ☐ Proportion of Treatment Clusters
- ☐ Total Number of Clusters
- ☒ Number of Individuals in Each Cluster

**Cost Information**

	Treatment	Control
New Group Fixed Cost	600	300
New Individual Cost	2	2
Total Cost Available		

**Desired Power or Width of Confidence Interval of Effect Size**

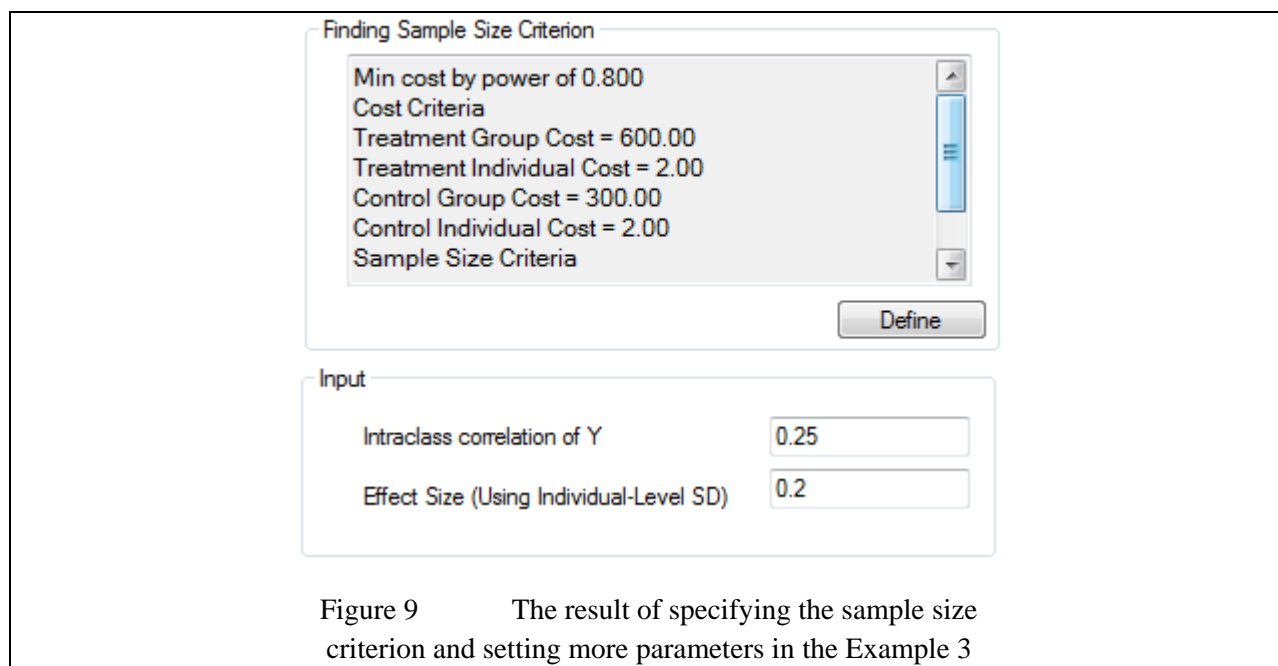
- ☒ Power
- ☐ Width of CI of Effect Size

Confidence Level

Degree of Certainty with level of

OK Cancel

Figure 8 Specifying a criterion for sample sizes estimation in the Example 3

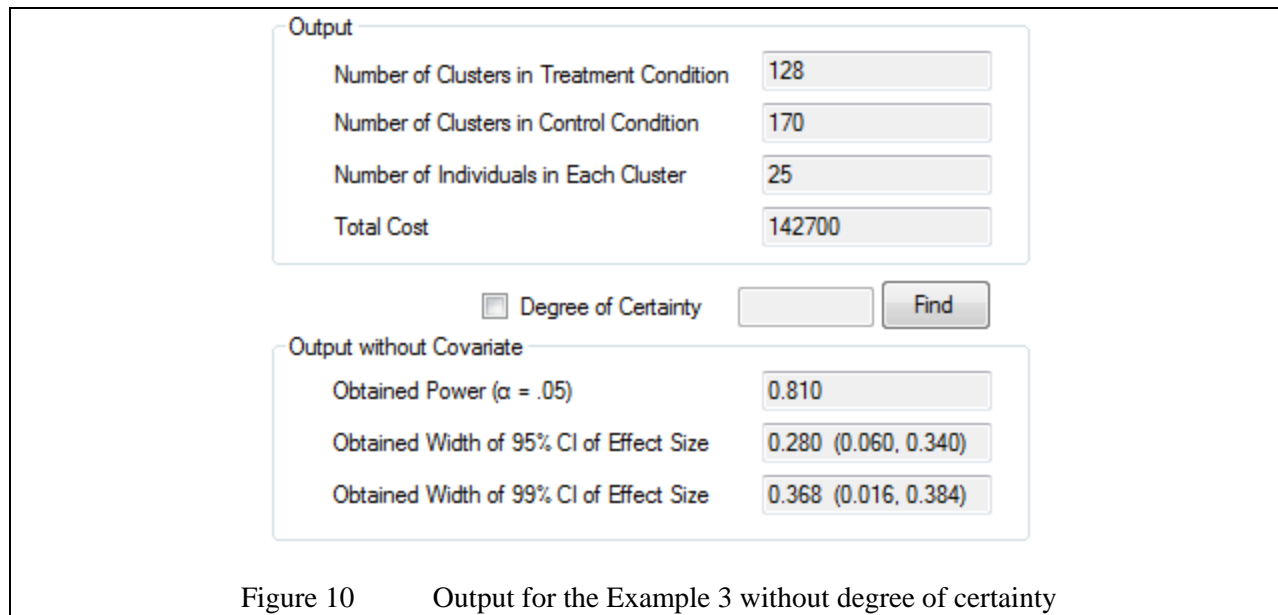


The screenshot shows the 'Finding Sample Size Criterion' window. The 'Min cost by power of 0.800' criterion is selected. The 'Cost Criteria' section lists: Treatment Group Cost = 600.00, Treatment Individual Cost = 2.00, Control Group Cost = 300.00, and Control Individual Cost = 2.00. The 'Sample Size Criteria' section is empty. A 'Define' button is at the bottom right. Below this is the 'Input' section with two fields: 'Intraclass correlation of Y' set to 0.25 and 'Effect Size (Using Individual-Level SD)' set to 0.2.

Figure 9 The result of specifying the sample size criterion and setting more parameters in the Example 3

### Outputs

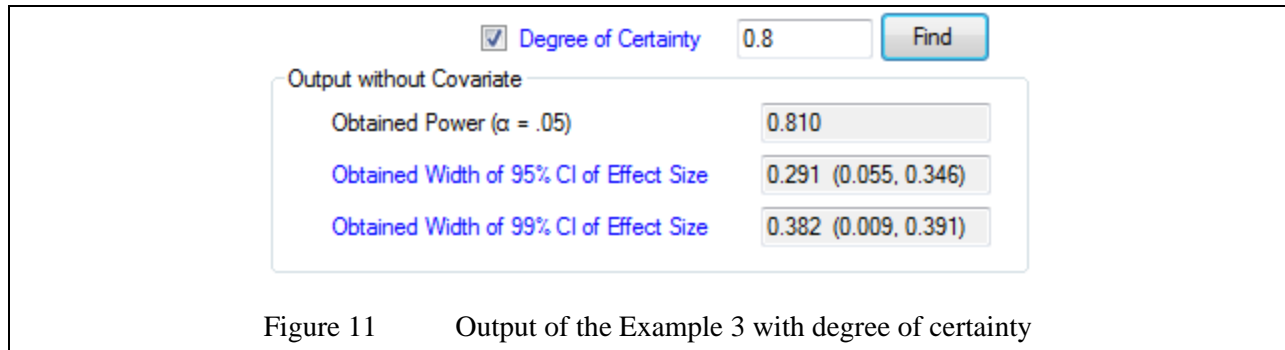
Based on the specification, as shown in Figure 10, the number of experiment groups is 128, the number of control groups is 170, and the number of individuals in each group is 25. The total number of individuals is  $(128 + 170) \times 25 = 7450$ . The total cost is \$142,700.



The screenshot shows the 'Output' window. It displays the following results: Number of Clusters in Treatment Condition = 128, Number of Clusters in Control Condition = 170, Number of Individuals in Each Cluster = 25, and Total Cost = 142700. Below this is a checkbox for 'Degree of Certainty' which is unchecked, and a 'Find' button. At the bottom is the 'Output without Covariate' section, which shows: Obtained Power ( $\alpha = .05$ ) = 0.810, Obtained Width of 95% CI of Effect Size = 0.280 (0.060, 0.340), and Obtained Width of 99% CI of Effect Size = 0.368 (0.016, 0.384).

Figure 10 Output for the Example 3 without degree of certainty

Based on the sample size combination result, the estimated power is 0.81. The average widths of 95% and 99% CI of ES are 0.28 (range from 0.060 to 0.340) and 0.368 (range from 0.016 to 0.384), respectively. Let specify the **Degree of Certainty** as 0.8. The result is shown in Figure 11. If researchers collect the data with the sample size combination, the 80% of width of 95% and 99% CI of ES of the obtained



sample will be less than 0.291 (corresponding to the range from 0.055 to 0.346) and 0.382 (corresponding to the range from 0.009 to 0.391).

### Example 4

Based on the previous example, the researchers would like to find CI of ES of the difference between the achievement averages of the new teaching method and the traditional method. They would like to find the result has 80% chances to get a sample size combination with the width 95% CI of ES less than 0.4 (i.e., margin of error of  $\pm 0.2$ ). In other words, the degree of certainty is 0.80 for the width of 95% CI of ES of 0.4. They think that the parametric assumptions of the CRD model probably cannot be achieved. Thus, they would like to have the same number of classrooms for the new teaching method and the traditional method. That is, the proportion of experiment clusters is 0.5.

### Inputs

As the example 3, the **A Priori** tab is chosen. Then, click **Define** below the **Find Sample Size Criterion** group box. Choose the second criterion, **Minimize the Cost with a Specified Power or Width**. As shown in Figure 12, the Pre-Specified Sample Size is modified. Like the previous example, the **Number of Individuals in Each Cluster** is checked and specified as 25. In this example, the **Proportion of Treatment Clusters** is also checked and specified as 0.5. The **Cost Information** is the same as the previous example. In the last group box, instead of choosing **Power**, the **Width of CI of Effect Size** will be chosen. The desired width will be put in the box in the **Width of CI of Effect Size** line as 0.40. In **Confidence Level**, the 95% and 99% confidence intervals are available. In this example, I will choose 95%. Researchers can specify whether they will use the **Degree of Certainty** function. If they do not use it, they will choose **No** in the drop-down list. If they use the degree of certainty, as in this example, they will choose **Yes (.50 to .99)**. Note that the a priori Monte Carlo simulation is needed for finding sample size combination based on the degree of certainty. After choosing **Yes (.50 to .99)**, the degree of certainty level should be specified. The possible value ranges from .50 to .99. In this example, the .80 is put in the box. Then, click **OK**. The specification is showed in the **Find Sample Size Criterion** textbox. Next, as the previous example, the **Effect Size (Using Individual-Level SD)** and the **Intraclass Correlation of Y** are specified as 0.20 and 0.25, respectively. Finally, click **Calculate**.

### Outputs

Based on the specification, as shown in Figure 13, the number of experiment groups is 78 and the number of control groups is 79. Even though I have already specified the proportion of treatment groups as 0.5, the result of total number of groups is odd. Thus, the program arbitrarily rounds down one condition and rounds up another condition. The researchers may round up both conditions and collect 79 treatment

**Criteria for Finding Sample Size**

**Input Criterion**

- ☐ Minimize the Total Number of Individuals with a Specified Power or Width
- ☒ Minimize the Cost with a Specified Power or Width
- ☐ Maximize the Power or Minimize the Width with a Limited Budget

**Pre-Defined Sample Size**

Do you wish to specify the sample size results in advance?

- ☒ Proportion of Treatment Clusters: 0.5
- ☐ Total Number of Clusters:
- ☒ Number of Individuals in Each Cluster: 25

**Cost Information**

	Treatment	Control
New Group Fixed Cost	600	300
New Individual Cost	2	2
Total Cost Available		

**Desired Power or Width of Confidence Interval of Effect Size**

- ☐ Power:
- ☒ Width of CI of Effect Size: 0.40
- Confidence Level: 95%
- Degree of Certainty: Yes (.50 to .99) with level of 0.80

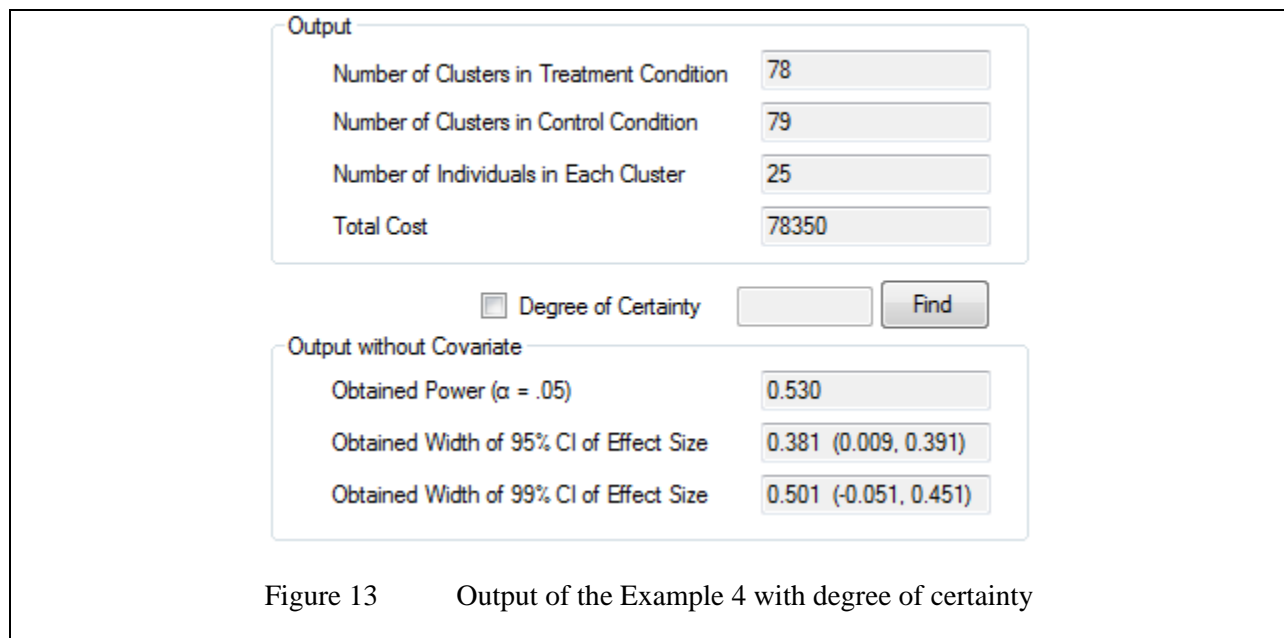
OK Cancel

Figure 12 Specifying a criterion for sample sizes estimation in the Example 4

groups and 79 control groups. As specified, the number of individuals in each group is 25. The total number of individuals is  $(78 + 79) \times 25 = 3925$ . The total cost is \$78,350.

Based on the sample size combination result, the estimated power is 0.53. The average widths of 95% and 99% CI of ES are 0.381 (range from 0.009 to 0.391) and 0.501 (range from -0.051 to 0.451), respectively. Let specify the **Degree of Certainty** as 0.8. The result is shown in Figure 11. If researchers collect the data with the sample size combination, the 80% of width of 95% and 99% CI of ES of the obtained sample will be less than 0.399 (corresponding to the range from 0.001 to 0.400) and 0.524





Output

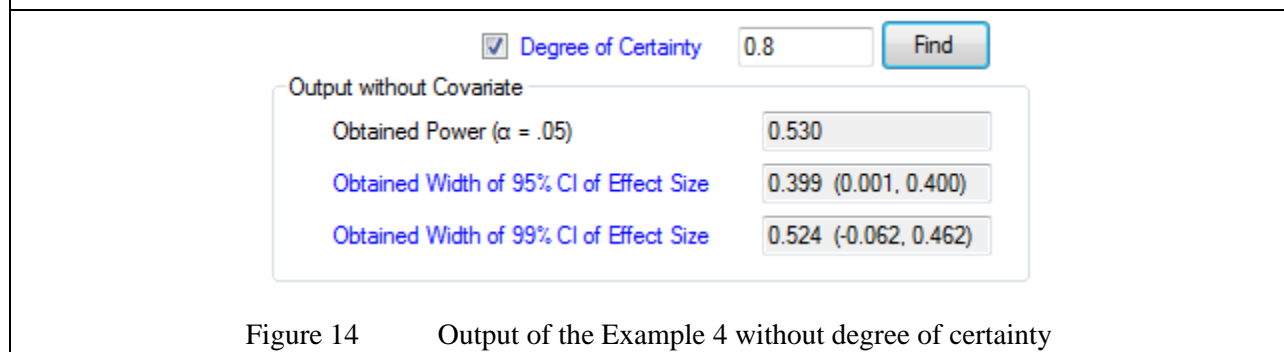
Number of Clusters in Treatment Condition	78
Number of Clusters in Control Condition	79
Number of Individuals in Each Cluster	25
Total Cost	78350

☐ Degree of Certainty

Output without Covariate

Obtained Power ( $\alpha = .05$ )	0.530
Obtained Width of 95% CI of Effect Size	0.381 (0.009, 0.391)
Obtained Width of 99% CI of Effect Size	0.501 (-0.051, 0.451)

Figure 13      Output of the Example 4 with degree of certainty



☒ Degree of Certainty 0.8

Output without Covariate

Obtained Power ( $\alpha = .05$ )	0.530
Obtained Width of 95% CI of Effect Size	0.399 (0.001, 0.400)
Obtained Width of 99% CI of Effect Size	0.524 (-0.062, 0.462)

Figure 14      Output of the Example 4 without degree of certainty

(corresponding to the range from -0.062 to 0.462). As specified, the 80% degree of certainty of the 95% width of CI of ES is 0.399.

### Example 5

Educational researchers would like to examine whether they should implement their parents-teachers relationship encouragement program. Thus, the researchers would like to test the program by randomly assigning some schools to the experiment condition and other schools to the control condition. They do nothing for the schools in the control condition. The dependent variable is the conduct problems score based on BASC teaching rating scale. The researchers think that the effective program should have the effect size at least 0.2. The intraclass correlation is 0.05, based on the Hedges and Hedberg's (2007) guideline.

The researchers consider perceived positive school climate (i.e., students' perception whether the staffs in schools care the students as individuals) as a covariate. Sameroff, Peck, & Eccles (2004) found that this covariate correlated with conduct problems about .43. The participants in this study involved many schools in a same county.

The costs of collecting data from individuals within treatment and control schools are \$30 and \$2, respectively. The treatment individual cost is higher than the control individual cost because some parts of

the treatment involves with parents-teachers discussion individually. Even though it is not the part of the analysis, it should include in the research plan because the cost increases as the researchers add a new individual. The costs of collecting data from a new school on the treatment and control conditions are \$500 and \$50, respectively. However, the researchers have a limited budget of \$50,000. Also, the researchers would like to know how much usefulness of including a covariate in the model.

### Inputs

First, click the **A Priori** tab on the top left of the window. Next, click the **Define** button below the **Find Sample Size Criterion** group box to specify a criterion for sample sizes estimation. The new pop-up window will appear as Figure 15.

Because the researchers would like to find the optimal design (i.e., the sample size combination with the highest level of power or the lowest level of the width of CI of ES for a limited budget, the third criterion (**Maximize the Power or Minimize the Width with a Limited Budget**) in the **Input Criterion** group box will be chosen. In the **Pre-Defined Sample Size** box, researchers will consider about pre-specified characteristics of the result of sample size combination. In this example, the researchers do not constrain the number of groups, number of individuals in each group, or proportion of treatment groups in advance. Next, **Cost Information** are specified. The treatment group cost is 500. The control group cost is 50. The individual costs for both treatment and control conditions are 30 and 2, respectively. The total available cost is 50000. The last group is not available in the third criterion. After everything is set, click **OK**.

After that, the researchers need to let the program know that the model includes a covariate. Thus, as shown in Figure 16, check **Activate Covariate**. Because the covariate varies in both levels, make sure that the **Cluster-level Covariate** and **Individual-level Covariate** checkboxes are clear. Also, the researchers would like to compare a model with and without a covariate, so the researchers can check **Also Activate No Covariate Simulation** to make the program analyze the summary for the power, width of 95% and 99% CI of ES of the model with and without a covariate. It allows the researchers to see the benefits of including a covariate. When the covariate varies in both levels, the last checkbox, **Covariate effects on both levels have different signs (+/-)**, is available. This checkbox is used when a researcher expects that the effects of the covariate on the dependent variable in the individual and cluster level are in different directions. For example, researchers would like to see the effect of level of discipline on intention to leave. The data are collected from incumbents from different organizations. In the incumbent level, perceived level of discipline may increase the intention to leave. However, in the organization level, the lenient organization (i.e., low level of the average perceived level of discipline across the incumbents in an organization) tends to have more average intention to leave than the harsh organization (i.e, high level of the average perceived level of discipline). However, in this example, the covariate should not have the effects in different direction on both levels. Thus, the **Covariate effects on both level have different signs (+/-)** should be clear.

As shown in Figure 17, for the dependent variable, the **Intraclass Correlation of Y** is 0.05 and the **Effect Size (Using Individual-Level SD)** is 0.20. For the covariate, the researchers know the correlation between the covariate and the dependent variable. However, the correlation of .43 ( $R^2 = .1849$ ) is computed from all students regardless of their membership in the schools. The correlation coefficient represents covariations of scores across both student and school levels. The researchers may have no idea about the magnitude of the covariate effect (i.e., the correlation coefficient) on the dependent variable in

both levels. It may be a good idea to explore various magnitudes of effects, such as higher, equal, and lower than the obtained correlations in both levels. For illustration, I assume that the correlations between perceived school climate and conduct problems within a school and between schools are equal to the

Criteria for Finding Sample Size

Input Criterion

☐ Minimize the Total Number of Individuals with a Specified Power or Width

☐ Minimize the Cost with a Specified Power or Width

☒ Maximize the Power or Minimize the Width with a Limited Budget

Pre-Defined Sample Size

Do you wish to specify the sample size results in advance?

☐ Proportion of Treatment Clusters

☐ Total Number of Clusters

☐ Number of Individuals in Each Cluster

Cost Information

	Treatment	Control
New Group Fixed Cost	500	50
New Individual Cost	30	2
Total Cost Available	50000	

Desired Power or Width of Confidence Interval of Effect Size

☐ Power

☐ Width of CI of Effect Size

Confidence Level

Degree of Certainty

with level of

OK

Cancel

Figure 15      Specifying a criterion for sample sizes estimation in the Example 5

☒ Activate Covariate

☐ Cluster-level Covariate

☒ Use A Priori Monte Carlo Simulation

☒ Also Activate No Covariate Simulation

☐ Individual-level Covariate

☐ Covariate effects on both levels have different signs (+/-).

Figure 16      Checkbox characteristics for the Example 5

Input	
Intraclass correlation of Y	0.05
Effect Size (Using Individual-Level SD)	0.20

Input Covariate	
Intraclass correlation of the covariate	0.05
Cluster Level Error Variance Explained	0.1849
Individual Level Error Variance Explained	0.1849

Figure 17 Inputs for the Dependent Variable and the Covariate in the Example 5

obtained correlation. Thus, the proportions of **Cluster Level Error Variance Explained** and **Individual Level Error Variance Explained** are both 0.1849.

The researchers have no idea about the intraclass correlation of the perceived positive school climate as well. I suggested that, for sample size estimations, the researchers may impute various numbers of intraclass correlations and pick one which have highest sample sizes, such as impute 0.05, 0.15, and 0.25, for intraclass correlations. For illustration, I enter 0.05 for the **Intraclass correlation of the covariate**.

### Outputs

Based on the specification, as shown in Figure 18, the number of experiment groups is 38 and the number of control groups is 133. The number of individuals in each group is 17. The total number of individuals is  $(38 + 133) \times 17 = 2907$ . The total cost is \$49,522 within the budget of \$50,000.

Based on the sample size combination result, the model with the covariate provides the power level of 0.948, the average width of 95% CI of ES of 0.220 (range from 0.090 to 0.310), and the average width of 99% CI of ES of 0.289 (range from 0.056 to 0.345). The model without a covariate provides the power level of 0.904, the average width of 95% CI of ES of 0.239 (range from 0.080 to 0.320), and the average width of 99% CI of ES of 0.314 (range from 0.043 to 0.357). The differences between the model with and without the covariate are not large and the model without covariate has the large power and the small widths of CI of ES. Therefore, the researchers may exclude the covariate from the model.

Again, the researchers may consider the **Degree of Certainty**. If researchers collect the data with the sample size combination with the covariate, the 80% of width of 95% and 99% CI of ES of the obtained sample will be less than 0.239 (corresponding to the range from 0.081 to 0.320) and 0.314 (corresponding to the range from 0.043 to 0.357). However, in the model without the covariate, the 80% of width of 95% and 99% CI of ES of the obtained sample will be less than 0.259 (corresponding to the range from 0.071 to 0.330) and 0.341 (corresponding to the range from 0.030 to 0.371). The differences between the degrees of certainty of both models are not large.

Output

Number of Clusters in Treatment Condition	38
Number of Clusters in Control Condition	133
Number of Individuals in Each Cluster	17
Total Cost	49552

☐ Degree of Certainty  Find

Output without Covariate

Obtained Power ( $\alpha = .05$ )	0.904
Obtained Width of 95% CI of Effect Size	0.239 (0.080, 0.320)
Obtained Width of 99% CI of Effect Size	0.314 (0.043, 0.357)

Output with Covariate

Obtained Power ( $\alpha = .05$ )	0.948
Obtained Width of 95% CI of Effect Size	0.220 (0.090, 0.310)
Obtained Width of 99% CI of Effect Size	0.289 (0.056, 0.345)

Figure 18 Output of Example 5 without degree of certainty

☒ Degree of Certainty  Find

Output without Covariate

Obtained Power ( $\alpha = .05$ )	0.904
Obtained Width of 95% CI of Effect Size	0.259 (0.071, 0.330)
Obtained Width of 99% CI of Effect Size	0.341 (0.030, 0.371)

Output with Covariate

Obtained Power ( $\alpha = .05$ )	0.948
Obtained Width of 95% CI of Effect Size	0.239 (0.081, 0.320)
Obtained Width of 99% CI of Effect Size	0.314 (0.043, 0.357)

Figure 19 Output of the Example 5 with degree of certainty

## Multiple Covariates

Basically, researchers need to combine multiple covariates to a single covariate to include in the program. As in the program, the researchers have three different types of a combination of multiple covariates.

First, if all covariates are cluster-level, researchers can specify the aggregated covariate as cluster-level covariate. Then, researchers need to specify only the proportion of cluster error variance explained.

Second, if all covariates are individual-level, researchers can specify the aggregated covariate as individual-level covariate. Then, researchers need to specify only the proportion of individual error variance explained.

Finally, in other cases, the intraclass correlation, the proportion of cluster error variance explained, and the proportion of individual error variance explained are needed to be speculated. Researchers might be able to speculate the proportions of cluster and individual error variances explained. However, the intraclass correlation of the aggregated covariate is very hard to estimate. It depends on intraclass correlations of each variable and the regression coefficients of the dependent variable on each covariate in both levels. Although all covariates have the same intraclass correlations, the aggregated covariate may not equal the intraclass correlation of each variable. In this case, I recommend that researcher use PINT instead of PAWS-CRD.

## Reference

- Byrk, A. S. & Raudenbush, S. W. (2002). *Hierarchical linear model: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Cheung, M. W. –L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling*, 16, 267-294.
- Cochran, W. J. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121-138.
- Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 341-370.
- Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60-87.
- Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson.
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19-24.
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11, 363-385.
- King, M., Davidson, O., Taylor, F., Haines, A., Sharp, D., & Turner, R. (2002). Effectiveness of teaching general practitioners skills in brief cognitive behavior therapy to treat patients with depression: randomized control trial. *British Medical Journal*, 324, 947-950.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537-563.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9, 599-620.
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.
- Sameroff, A. J., Peck, S. C., & Eccles, J. S. (2004). Changing ecological determinants of conduct problems from early adolescence to early adulthood. *Development and psychopathology*, 16, 873-896.
- Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 237-259.

- Thompson, B. (2002). What future of quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31, 25-32.
- Wilkinson, L. & the Task Force of Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594-604.