CHAPTER I

THE PROBLEM AND ITS BACKGROUND

Statistical power, the probability of finding significant results for a given effect in a population, rises when sample size increases. The selection of the optimal sample size involves a balancing of concerns about statistical power and about budgeting and other resources. A common practice that balances these concerns is to find a sample size that ensures a power of .80 or .90. Several free and high quality programs for power calculation in simple designs are available (e.g., *G*Power*; Faul, Erdfelder, Lang, & Buchner, 2007). Moreover, free and high quality programs for more complex designs, such as the cluster randomized design, are also available. These programs, however, can be improved upon because each of them has different shortcomings and limitations. This study describes the parameters and theoretical underpinnings of such an attempt to make these improvements: a program called Power Analysis and Width of confidence interval for Sample sizes estimation (PAWS), for power calculation in the case of the cluster randomized design.

The cluster randomized design (CRD) is a design in which the unit of random assignment to different conditions is groups of objects, such as people participating in experiments about group psychotherapy, classrooms, schools, departments, or organizations. One of the complexities associated with the statistical analysis of data with

this structure is that there are two types of sample size: number of clusters (i.e., groups) and cluster size. Because different combinations of these two types of sample sizes may provide the same power, selecting a particular combination depends on the researcher's goals, priorities, and budget. PAWS allows researchers to select a combination of the number of clusters and cluster size depending on whether the goal is to achieve a particular threshold of power while minimizing cost or to maximize power given a fixed budget. PAWS can find the optimal sample size for a specified width of the confidence interval (CI) of an effect size (ES) as well.

Because power analysis is a well-known procedure, I will not discuss it in detail. Instead, I will focus on providing the background knowledge necessary to understand issues related to the accuracy in parameter estimation, given by a width of confidence intervals of effect size (CI of ES). Next, I will explain the relationship between significance testing and CI of ES. After that, I will explain how to find the CI of Cohen's *d* for the difference between independent means, which is the simplest case of CI of ES. Next, the similarities and differences in finding sample size based on power and desired accuracy in effect size estimation will be explained. The sample size based on both ways for comparing two independent means will be discussed. Finally, I will illustrate these principles using CRD.

Rationale of Confidence Interval of Effect Size

During the 1990's, significance testing was challenged and criticized with increasing frequency by many researchers (Cohen, 1994; Nickerson, 2000). Wilkinson and the Task Force of Statistical Inference (1999) encouraged researchers to use ES and interval estimations of ES. One of the benefits of reporting the CI of ES is that not only

does it provide all of the information provided by significance testing but also adds additional information that can be very useful to researchers. It is true that some research questions are concerned solely with the direction of the effect (Maxwell, Kelley, & Rausch, 2008). Many research questions also involve estimates of the magnitude of the effect. In these studies, the CI of ES offers considerable advantages over significance testing.

Significance Testing and Confidence Intervals of Effect Sizes

Significance testing answers the question of whether the effect really exists in a population. The CI of ES provides information on both the magnitude of the effect and also the precision of the effect size estimate. There are at least five advantages of using CI of ES over using traditional significance testing when analyzing data.

First, the CI of ES gives more direct and more useful information about ES than does traditional significance testing. Although some researchers may use *p*-values to loosely infer ES (e.g., distinguishing between *p*-values of .05, .01, and .001 or declaring an effect "marginally significant," "significant," or "highly significant"), this practice can lead to erroneous conclusions. It is true that the smaller the *p*-value, the greater the distance from the boundary of CI of ES to the null hypothesis value. This relationship, however, is nonlinear. Furthermore, the *p*-value is determined by the sample size, such that a very small effect can have a very small *p*-value if the sample size is large. Thus, ES estimates and CI of ES provide much more direct and easily understood information about the magnitude of effects than do the *p*-value.

Second, the CI of ES reveals the precision of the point estimation (Wilkinson et al., 1999). ES statistics are point estimates. Although researchers may obtain high values

of ES, the CI may be large and thus may fluctuate widely in replication studies. When a CI of ES is narrow, researchers can be confident that the point estimation of the ES is stable. In contrast, significance testing provides a significance value, which only helps the researcher decide if the effect size parameter is different from zero.

Third, the CI of ES can be used for a surprising purpose: accepting the null hypothesis (Kelley & Rausch, 2006). As most statistics students have been warned repeatedly by their instructors, significance testing cannot be used for that purpose. Just because a researcher fails to reject the null hypothesis, it is not the case that the null hypothesis has been proven. In contrast, if the ES is very small and the CI of ES is very narrow, then one can argue that whatever the true size of the effect, it is of negligible importance.

Fourth, the CI of ES is useful for integrating the results of many research studies, as is done in meta-analysis (Cumming & Finch, 2001). Researchers may use the CI of ES to aggregate results across studies and compare groups of studies. Thompson (2002) illustrated how to combine the CI of ES across studies to a single CI of ES. He showed that 11 nonsignificant studies can combine to a single CI of ES which did not include zero. Combining effect sizes across different studies makes the estimation of the ES more precise. The aggregation of the results makes the estimation of the ES have a narrower CI.

Finally, the CI of ES can be used to test a broader set of hypotheses than is typically conducted with null hypothesis testing. A researcher might not be interested in whether an effect is larger than zero. Rather, the researcher might wish to know if one ES is significantly larger than some other ES. Suppose a standard treatment has an effect size

of 0.5. A new treatment may have an estimated effect size of 0.6. Without a CI of ES, it is difficult to make the case that the new treatment is superior to the standard treatment unless the treatments have been compared in a head-to-head clinical outcome trial. If the CI of ES is sufficiently small, a researcher can reasonably infer that the new treatment is better, even though the two treatments were never compared directly.

Reporting Accuracy in Parameter Estimation

Alhija and Levy (2009) estimate that only half of educational research studies reported ES statistics during 2003-2004. Even when ES statistics were reported, it was quite rare for the CI of ES statistics to accompany them. Ignorance of the existence of CI of ES is probably the main reason that researchers report them infrequently. However, Cohen (1994) suggested that one reason for the unpopularity of CI of ES was that the CI's from most research studies were "embarrassing large," which may reduce the likelihood of the studies being published (Maxwell et al., 2008). The widespread failure to report CI of ES is unfortunate because CI of ES potentially could provide researchers with important information and can help guide sample size selection (Kelley & Rausch, 2006; Maxwell et al., 2008).

Power and CI of ES in Two Independent Means Difference

One of the most important issues related to CI of ES has to do with setting the desired accuracy in parameter estimation, which is determined by the width of CI of ES. The smaller the CI of ES, the more precise the estimation of the parameter is. This concept can most easily be illustrated with the simplest experimental design: comparing two independent means.

Many different procedures have been developed to estimate the CI of ES of the difference of two independent means. These options are based on different approaches: parametric or nonparametric, accurate or approximate, and different versions of Cohen's *d* statistic, such as one which is based on means or trimmed means. When the assumptions of parametric statistics hold, the standard parametric statistics comparing independent means are the most efficient, which yields the narrowest standard error and a CI that is more likely to contain the true population parameter (Bonett, 2008; Keselman, Algina, Lix, Wilcox, & Deering, 2008). When parametric assumptions (e.g., the normality assumption or the homogeneity of variance assumption) are violated, however, other methods, such as the bootstrap CI (Efron & Tibshirani, 1994), produce CI that are more likely to contain the true population parameter.

Because the main purpose of this article is to estimate sample size, I will assume that the parametric assumptions of estimating power or CI of ES hold. When researchers are attempting to estimate how large a sample is needed for a particular study, they often do not know whether the variables they intend to measure are normally distributed. Likewise, researchers often do not know by how much the variance of the dependent variables will differ across groups. Therefore, I will rely on the standard assumptions of parametric tests and provide the sample size. The standard parametric statistics method for comparing two independent means is based on the standard independent *t*-test. The formula is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1}$$

where $\bar{X}_1$ and $\bar{X}_2$ are group means of condition 1 and 2, respectively, n_1 and n_2 are sample size of group 1 and 2, and s_p is pooled standard deviation from both groups. The standard error of this method is calculated by pooled standard deviation, which is the square root of the weighted mean of variance of two groups, assuming homogeneity of variance. The distribution of *t* statistics, when there is no mean difference in population, will be a central *t* distribution in which the degrees of freedom are given by $n_1 + n_2 - 2$. To estimate power, the distribution of the alternative hypothesis must be known. Most of the time, researchers do not know by how much the population means differ but estimate the difference using Cohen's *d* (Cohen, 1988), which is computed in samples or populations by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p};\ \delta = \frac{\mu_1 - \mu_2}{\sigma_p}, \tag{2}$$

where δ is the parameter of Cohen's *d*, μ_1 and μ_2 are group means of condition 1 and 2 in population and σ_p are pooled standard deviation from both groups in population. Other symbols are defined in the Equation 1.

The *t* formula can be written with the standardized mean difference *d* as

$$t = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{3}$$

If researchers specify the ES in advance, they must be aware that Cohen's *d* depends on the sample means of both groups and the pooled standard deviation and that both the means and the pooled standard deviation are subject to sampling error. With the exception of the case in which the null hypothesis is true, the distribution of the *t* statistic is a noncentral *t* distribution instead of the central *t* distribution. The noncentral *t*

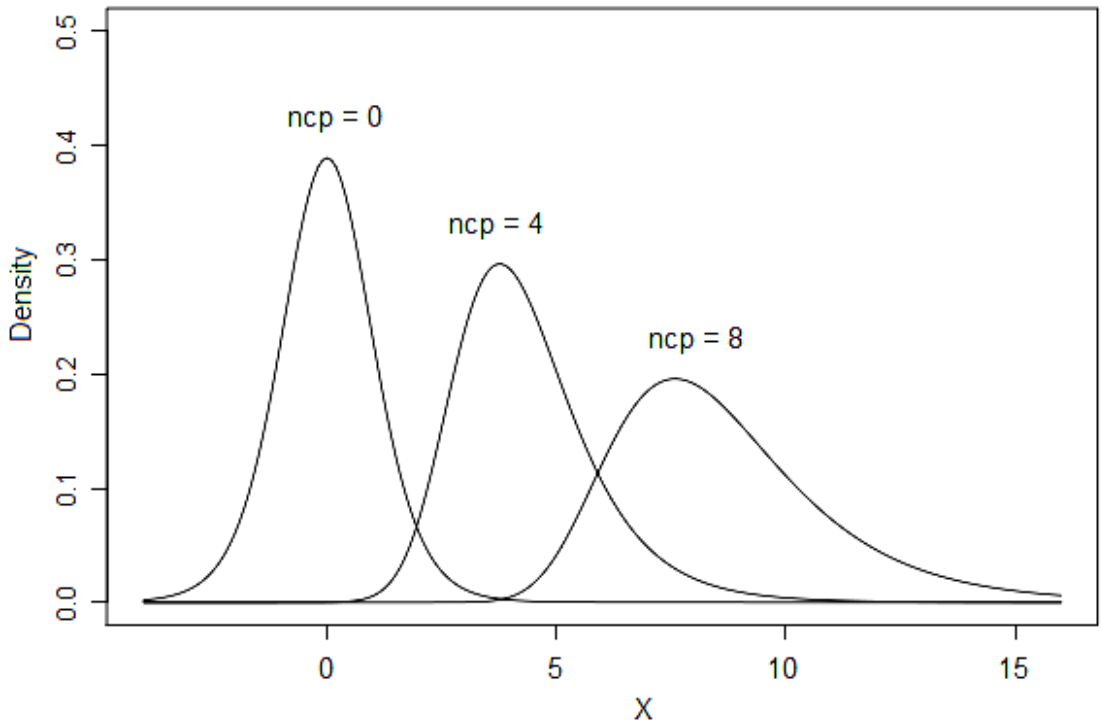

Figure 1. The noncentral *t* distribution is asymmetrical. The center of noncentral *t* distribution is called the noncentrality parameter (*ncp*). The larger the *ncp*, the more asymmetric the distribution is. The distribution that *ncp* = 0 is the central *t* distribution, which is symmetrical.

distribution depends not only on the degrees of freedom ($n_1 + n_2 - 2$) but also on the magnitude of the effect size. As shown in Figure 1, the noncentral *t* distribution will be nonsymmetrical and wider than the central *t* distribution when the degrees of freedom are low and the magnitude of effect size is high. The exact center of the noncentral *t* distribution, called the noncentrality parameter (Δ), is

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma_p}, \tag{4}$$

which is equal to Cohen's *d* ES in the independent-samples t-test. The statistical power of the independent-samples *t*-test is determined from the proportion of the noncentral *t* distribution that is greater than the critical value of the central *t* distribution when testing the null hypothesis.

To develop the idea of a CI of ES for the independent *t*-test, I will show how to build the CI of raw score means differences, which may be developed in two different ways (Cumming & Finch, 2001). The first method is to solve for the CI using the *t* statistic formula. Then, the formula for the CI for raw score means difference will be

$$\mathrm{CI}_{1-\alpha} \text{ of } \mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm \left(t_{df,\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right), \tag{5}$$

where α is significance level, $\mathrm{CI}_{1-\alpha}$ of $\mu_1 - \mu_2$ is confidence interval of raw score mean difference with confidence level of 1- α, $t_{df,\alpha}$ is critical *t* value which the degrees of freedom of *df* and 2-tailed significance level of α.

This method assumes that the distributions of *t* statistics for raw score differences are identical for the upper and lower bounds of the CI. The sampling distribution in the lower and upper bounds are identical because both statistics have a central *t* distribution. The reason they both have a central *t* distribution instead of a noncentral *t* distribution is that the null hypothesis is not specified in the CI of raw score mean differences. Therefore, the CI shows raw score difference parameters, null hypotheses values, which can draw the obtained raw score mean difference.

Another method will start with the mean difference parameter, $\mu_1 - \mu_2$. Next, two sampling distributions of $\bar{X}_1 - \bar{X}_2$ are drawn on the left and right hand side of obtained $\bar{X}_1 - \bar{X}_2$, which is the point estimate of $\mu_1 - \mu_2$. Then, the distribution of $\bar{X}_1 - \bar{X}_2$ on the left side is moved until the critical value on the right side equals obtained $\bar{X}_1 - \bar{X}_2$, as well as moving the distribution on right side until the critical value on the left side reaches obtained $\bar{X}_1 - \bar{X}_2$. The lower and upper bound of the CI is the center of the distribution of left and right sides. The whole process is shown in the top panel of Figure 2. This method can be used when the upper and lower bounds of the CI have either identical or different sampling distributions.

The CI of ES is related to the distribution of standardized mean differences, which have noncentral *t* distributions. Because the shape of the noncentral *t* distribution depends on the effect size (noncentrality parameter), the distance between the obtained ES and the lower bound is different from one between the obtained ES and the upper bound. The reason that the distribution of the ES has a noncentral *t* distribution is that the null hypothesis value must be specified for constructing ES, which is 0. It can be shown in formula as

$$\delta = \frac{(\mu_1 - \mu_2) - \text{Null hypotheses value}}{\sigma_p}, \tag{6}$$

where all symbols are defined in Equation 2. CI of ES shows the Cohen's *d* value in the alternative hypotheses which are likely to draw the obtained Cohen's *d*. The alternative hypotheses are distributed as noncentral *t*. Therefore, the second method is appropriate for constructing CI of ES. To develop the CI of a Cohen's *d*, two noncentral *t* distributions are drawn on the left and right hand side of the obtained *t*, which is

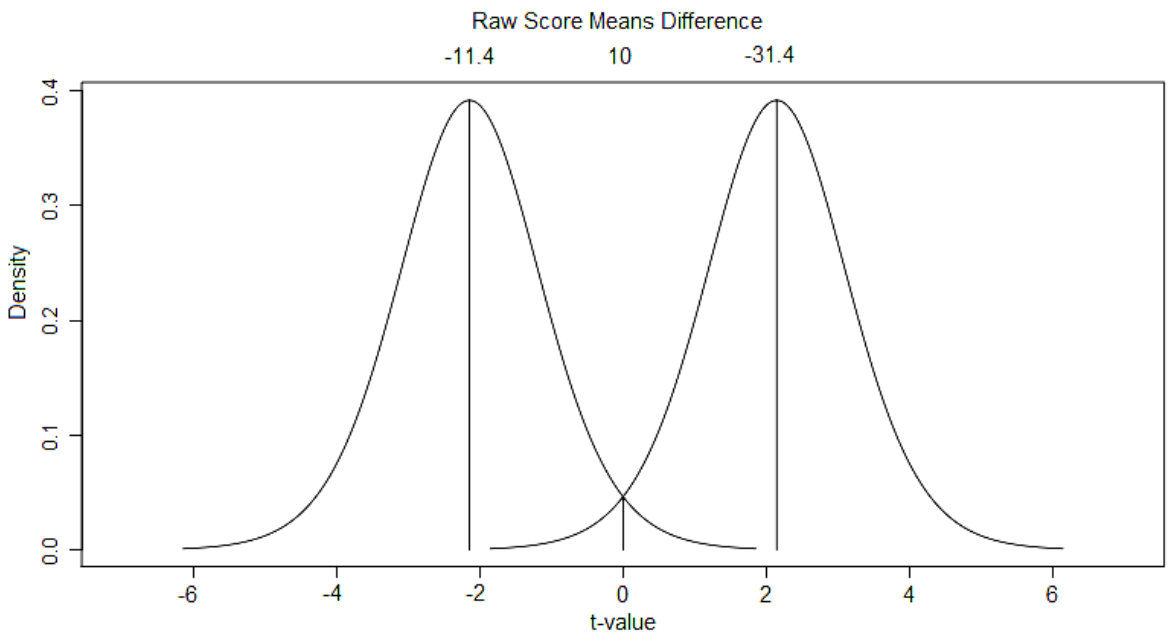

Figure 2. Two central *t* distributions are drawn on the left and right hand sides of the observed *t*. The critical values on the right (i.e., 97.5th percentile) and the left (i.e., 2.5th percentile) of the left and right distributions, respectively, are equal to the observed *t*. The lower and upper bounds of the CI of raw score means difference can be derived from the left and right distribution by *t* formula in the Equation 1.

transformed from the obtained Cohen's *d*. The noncentral *t* distribution on the left side is moved until the critical value on the right side (i.e., 97.5th percentile for 95% CI) equals the obtained *t*, as well as moving the noncentral *t* distribution on the right hand side until the critical value on the left side (i.e., 2.5th percentile for 95% CI) reaches the obtained *t*. The lower and upper bound of CI of ES is the center or noncentrality parameters of the noncentral *t* distribution on the left and right sides. The whole process is shown in the

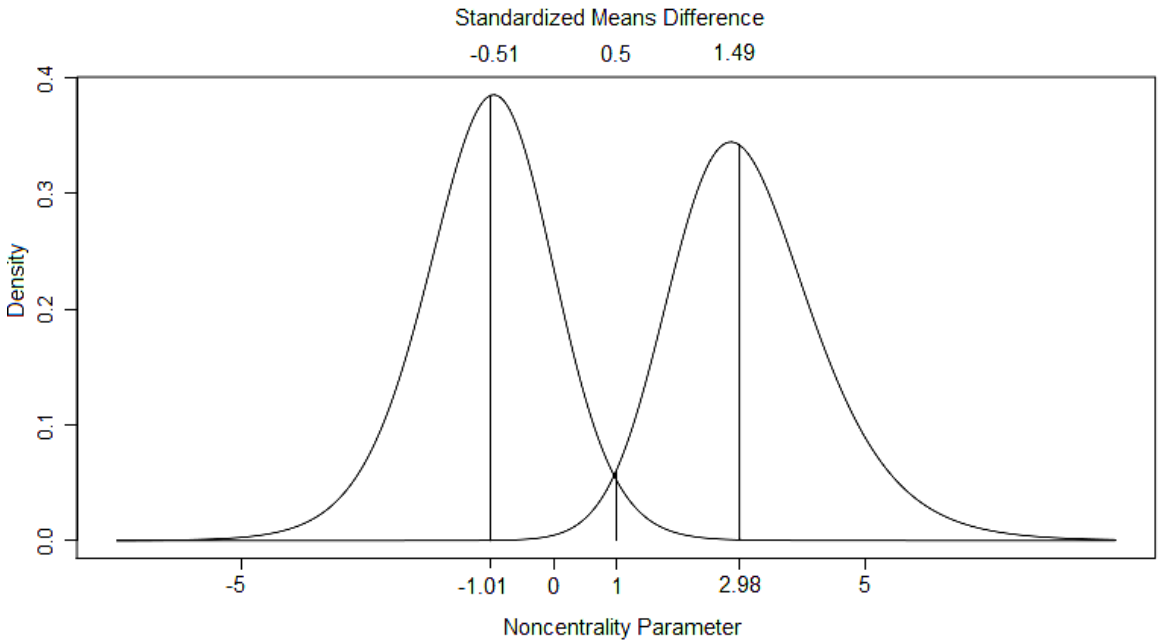

Figure 3. Two noncentral *t* distributions are drawn on the left and right hand sides of the observed *t*. The 97.5th and 2.5th percentiles of the left and right distributions, respectively, are equal to the observed *t*. The noncentrality parameters of the lower and upper distributions are transformed to the standardized means differences by the Equation 3.

Figure 3. If researchers change the null hypothesis value of raw score means difference, the ES statistic also changes and the CI of ES changes. The noncentral *t* distributions on the left and right sides change as well; therefore, the width of CI of ES changes.

The proportion of the area under the noncentral *t* distribution is often not substantially different from the analogous area under a normal curve, especially when the sample size is large and the ES is small. Although some researchers insist on using the more accurate noncentral *t* distribution (Cumming & Finch, 2001; Kelley & Rausch,

2006), others maintain that because estimates using the normal distribution are typically sufficiently close to estimates using the noncentral *t* distribution, the distinction is unimportant (Bonett, 2008, 2009). Bonett (2008) argued that, because computation of CI of ES is based on large sample sizes to make the CI as narrow as possible, the large sample approximation (i.e., using the normal curve) is sensible. Indeed, the noncentral *t* distribution rapidly converges to the normal distribution, as shown below in the sample size estimation of independent mean differences. To achieve the power of .80 in the medium ES, the choice between the noncentral *t* distribution and the normal distribution results in a negligible difference. Also, because researchers are not satisfied until the accuracy of estimation in the CI of ES is narrow (e.g., width of 0.5), the sample size required is quite large, as shown below. Thus, the distinction between the noncentral *t* distribution and the normal distribution is unimportant in this domain as well. In addition, a small difference in sample size estimation is not usually important because researchers typically plan to collect more data than the exact minimum sample size suggested by power analysis. Thus, although using the noncentral t distribution may result in a sample size recommendation that is slightly different from the sample size recommendation produced by procedures using the normal distribution, the actual difference in power or CI of ES that would result from using an "incorrect" sample size is usually negligible. In studies with a small sample size, however, researchers who would like to analyze CI of ES should use the noncentral *t* distribution, even though the CI is wide.

Sample Size Estimation for Two Independent Means

Sample size estimations for the purpose of ensuring adequate power often have different outcomes compared to sample size estimations for accuracy in parameter

estimation. When selecting a sample size that will ensure adequate power, researchers hope that they will obtain a statistically significant result. The purpose of sample size estimation for accuracy in parameter estimation, however, is to make sure that a sample statistic is a good estimate of a parameter. For example, researchers might wish to know how large the sample size must be to make sure that the Cohen's *d* effect size has a margin of error no larger than ±0.2. The different methods can require quite different sample sizes. I will demonstrate the sample size estimation process using the two independent means research design.

As shown above, to estimate power, researchers should know the noncentral *t* distribution of the ES statistic, the level of power they wish to have, and the critical value of the null hypothesis test. The procedure is an iterative process in which the sample size is increased until the probability of the area under the noncentral distribution above the critical value is equal to the specified power. Dalgaard (2008) provided the formula for estimating power when sample sizes in each group are equal as

$$n = 2 \times \left(\frac{z_{\alpha/2} + z_{1-\text{power}}}{\delta}\right)^2, \tag{7}$$

where z_x represents quantiles in the normal distribution, *n* is estimated sample size, and δ is the estimated parameter value of Cohen's *d*. This formula uses the normal distribution to approximate the noncentral *t* distribution. The estimated sample size is reasonably accurate when the *df* are over 20.

In finding a sample size for accuracy in parameter estimation, the process is similar to that of power analysis. Kelley and Rausch (2006) suggested that, as a first step, a researcher specifies a desired width of the CI of ES, the confidence level, and the

estimated population ES. These three numbers are used to estimate the starting value of the sample size (n_0) by

$$n_0 = \text{ceiling}\left[8\left(\frac{z_{(1-\alpha/2)}}{\omega}\right)^2\right], \tag{8}$$

where ω is a specified width of the CI of Cohen's *d* and ceiling[*x*] is the next largest integer function; in other words, the value is rounded up to an integer value rather than rounding to the closet integer. Notice that the starting value is estimated using the normal distribution. From there, the starting value is used for building the CI of Cohen's *d* by the procedure shown above. Two noncentral *t* distributions are drawn on the left and right hand side of the obtained *t*, which is transformed from the obtained Cohen's *d*. The noncentral *t* distribution on the left side is moved until the critical value on the right side (i.e., 97.5th percentile for 95% CI) equals the obtained *t*, as well as moving the noncentral *t* distribution on the right hand side until the critical value on the left side (i.e., 2.5th percentile for 95% CI) reaches the obtained *t*. The width of CI of Cohen's *d* is the distance between the lower and upper bounds. Next, as an iterative process, the sample size increases until the width of the CI of ES is equal to the desired width.

This width, however, is a population parameter, and as such, it is subject to sampling error such that the obtained width in 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 for 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 to reduce the probability of getting CI width larger than desired.

The probability that obtained samples have a CI of ES width less than or equal to the desired width is known as degree of certainty (Kelley, 2008; Kelley & Rausch, 2006).

Kelley and Rausch (2006) dealt with degree of certainty in finding sample sizes for accuracy in parameter estimation in two independent means by adjusting the ES parameter. They note that the farther the ES is from 0, the larger the width of the CI. If researchers use the ES (a.k.a. adjusted ES) that deviates from 0 more than the original ES to find the sample size, it will guarantee that the obtained sample size from the adjusted ES will produce the smaller width for the original ES. Therefore, the degree of certainty increases. In other words, the probability that the width from original ES is less than or equal to desired ES increases. The method to find the adjusted ES is shown in Kelley and Rausch (2006).

On the other hand, Bonett (2009) provided a simpler formula for estimating sample size for accuracy in parameter estimation for any linear contrasts to compare group means. The formula for two independent means difference is

$$n = (\delta^2 + 8)\left(\frac{z_{(1-\alpha/2)}}{\omega}\right)^2 + 1. \tag{9}$$

Bonett (2009) used a two-step approach to account for degree of certainty (γ). First, he provided an equation to find adjusted ES (δ_γ) by using the required sample size (n) from the Equation 9 using the original ES. The equation is

$$\delta_\gamma = \delta + z_\gamma \sqrt{\frac{\delta^2}{4(n-1)} + \frac{2}{n-1}}. \tag{10}$$

Next, the adjusted ES is used to find sample size again by Equation 9. Bonett (2009) said that the sample size estimation by this formula reproduced the sample size in Kelley and

Rausch (2006) table by a difference of 1. These formulas in Bonett (2009) are based on CI of ES formulas provided in Bonett (2008), which did not assume homogeneity of variance.

The results of power analysis and accuracy in parameter estimation are different because they are used for different purposes. According to Kelley and Rausch (2006), the sample size calculated for a given power is highly related to ES. If researchers wish to obtain a statistically significant difference from a small ES, they must have a large sample. For example, if researchers wish to have a power of .80 when the Cohen's *d* is 0.2, the estimated sample size is 788. The estimated sample size, however, is only 128 when the Cohen's *d* is 0.5. The ES and width of CI of ES relationship, however, is very small, so it may be considered a negligible relationship. On the other hand, the sample size calculated by accuracy in parameter estimation is highly related to the specified width. For example, if researchers wish to have the width of 95% CI of ES of 0.2 at the Cohen's *d* of 0.2 with the degree of certainty of .80, the estimated sample size is 774. The estimated sample size is 125, however, when the desired width is 0.5. The higher the required accuracy of the CI of ES, the larger the required sample size.

Extension to two-group CRD

The concept of accuracy can be applied to CRD as well. The formula from the independent t-test, however, cannot be used directly because the model must be account for both the number of clusters and cluster size. Even more complex, different numbers of clusters, cluster size, and proportion of treatment clusters can provide the same power and the same width of CI of ES. Therefore, researchers should pick a combination appropriate for their study, such as the lowest cost combination.

CRD has less power than the independent *t*-test when the total numbers of participants are equal (Hedges, 2007a); therefore, researchers should plan their studies carefully. Sometimes, increasing sample size is impossible. Researchers may find covariates to improve the precision of parameter estimates, which increases power and reduces the width of CI of ES (Raudenbush, Martinez, & Spybrook, 2007). I will show some studies that develop ways to estimate power in CRD and expand these ideas in some aspects.

Two-group CRD Model

The two-group CRD deals with a nested data structure. For the sake of clarity, I will refer to the higher level and lower level of the hierarchical data structure as the cluster level and the individual level, respectively. This two-group CRD model, however, can be applied to other situations, such as individual-measurements and school-classrooms. The treatment variable is applied in the higher level or cluster level, such as school-based intervention for school-individuals data structure. Treatment group membership may be either randomly assigned to clusters (e.g., two types of interventions) or not (e.g., nominal cluster characteristics such as the distinction between public and private schools).

In simple research designs, error typically refers to individual error (e). That is, error is the discrepancy between treatment condition means and individual data. In CRD, however, error is divided into two levels: cluster error (u) and individual error (e). Cluster error is the degree to which the cluster mean is not predicted by treatment condition. It can be conceptualized as the aggregated effect of the common experiences of all individuals within each cluster. For example, students in each classroom have similar

experiences: They have the same teacher and experience the same social climate in the class. When all students in a classroom have a good teacher, the cluster error in this class will be higher than the cluster error in other classrooms. In CRD, individual error represents the sum total of all unique effects that are uncorrelated with cluster error and the effects of the treatment variable. Examples of such effects might include individual difference variables such as cognitive abilities or personality. Therefore, each individual score can be divided into four components:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + u_j + e_{ij}, \tag{11}$$

which means that the score of individual *i* from cluster *j* (Y_{ij}) is the sum of the intercept (γ_0), the treatment condition effect (γ_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*. The intercept and treatment condition effect are interpreted differently, depending on how the treatment variable is coded. For usefulness in defining ES, I employed a 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 (γ_1) is the difference between the treatment condition and control condition means. The intercept (γ_0) is the control condition clusters mean. As shown in the Appendix, the variance of the difference between treatment and control condition means estimate is

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{\sigma_Y^2 + n\tau_Y}{nkp(1-p)}, \tag{12}$$

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

Cluster error variance and individual error variance (σ_Y^2) 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 total error variance is the sum of cluster and individual error variances. The intraclass correlation (ρ_Y) is the proportion of cluster error variance to total error variance. The greater the intraclass correlation, the more the similarity there is within clusters. The intraclass correlation can be defined as

$$\rho_Y = \frac{\tau_Y}{\tau_Y + \sigma_Y^2} \,. \tag{13}$$

The intraclass correlation is typically inversely related to the variance of the treatment condition difference estimate.

Two-group CRD Model with a Covariate

When a covariate is introduced for explaining individual scores, it is hoped that error variances are reduced. Cluster and individual error variance reduce in different degrees, however, 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. The between-cluster effect, however, 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,

$$Z_{ij} = \bar{Z}_j + \left(Z_{ij} - \bar{Z}_j\right), \tag{14}$$

where Z_{ij} is the covariate score of individual *i* from cluster *j* and is the average covariate score across individuals in cluster *j*. Including a covariate effect in the Equation 11, each dependent variable score can be divided as,

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + \gamma_B \bar{Z}_j + \gamma_W\left(Z_{ij} - \bar{Z}_j\right) + u_j + e_{ij}, \tag{15}$$

where γ_B is the group-level effect toward the dependent variable on the covariate, γ_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 Equations 11 and 14. This method of putting a covariate in the equation is also known as group-mean centering (Enders & Tofighi, 2007). The grand mean centering or no centering can be used, but the model specification to find power analysis is harder than group-mean centering. Therefore, I discuss only group-mean centering which divides the overall effect of a covariate into between- and within-group effects. Given the full range of options afforded by multilevel modeling, the individual-level covariate regression coefficient (γ_W) does not have to be equal across clusters. In this study, however, I 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 $\sigma^2_{(Y|Z)}$ 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

$$R_B^2 = \frac{\tau_Y - \tau_{(Y|Z)}}{\tau_Y} \tag{16}$$

$$R_W^2 = \frac{\sigma_Y^2 - \sigma_{(Y|Z)}^2}{\sigma_Y^2}, \tag{17}$$

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 (τ_Z) and individual-level (σ_Z^2). The intraclass correlation of the covariate (ρ_Z) is

$$\rho_Z = \frac{\tau_Z}{\tau_Z + \sigma_Z^2}. \tag{18}$$

As shown in the Appendix, the link between indices of proportions of variance explained and the covariate's regression coefficient are

$$\gamma_B^2 = R_B^2 \frac{\tau_Y}{\tau_Z} \tag{19}$$

$$\gamma_W^2 = R_W^2 \frac{\sigma_Y^2}{\sigma_Z^2}, \tag{20}$$

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

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{\sigma_{(Y|Z)}^2 + n\tau_{(Y|Z)}}{nkp(1-p)}. \tag{21}$$

A covariate, however, 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 Sizes in CRD Model

As shown above, Cohen's *d* or the standardized mean difference is the most popular ES statistics when the experimental design involves the comparison of two means. Cohen's *d*, however, 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, 2007b). 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. When most of the 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, however, the overall standard deviation is appropriate. As CRD is 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

$$d_W = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_Y} = \frac{\gamma_1}{\sigma_Y}, \tag{22}$$

where γ_1 is the regression coefficient when the grouping variable was coded as dummy variable (0 or 1) in Equations 11 and 15 and where σ_Y is the individual-level standard deviation.

Confidence Intervals of Effect Sizes

The approach for computing CI of ES by comparing two sample means cannot be easily generalized to the CRD model. In the independent samples *t*-test, the noncentrality parameter directly links with ES. In the two-group CRD model, however, the noncentrality parameter is a function of ES and intraclass correlation. Hedges (2007b) showed that the noncentral parameter in this distribution is

$$\Delta = \delta_W \cdot \sqrt{\frac{k_T k_C n}{k_T + k_C} \cdot \frac{1 - \rho_Y}{1 + (n-1)\rho_Y}}, \tag{23}$$

where k_T is the number of treatment groups, k_C is the number of control groups, *n* is number of individuals in each group, assuming that *n* is equal in both treatment and control conditions, ρ_Y is intraclass correlation of the dependent variable, and δ_W is parameter ES using the individual-level standard deviation. There are *n* – *k* degrees of freedom for the noncentrality distribution. Because the intraclass correlation is also a random variable (i.e., the intraclass correlation for different samples of clusters fluctuates), the CI of ES cannot be transformed from the lower and upper bounds of the noncentrality parameter.

Fortunately, there are at least two options available for computing CI of ES. I will show only the CI of ES using individual-level standard deviation. First, Hedges (2007b) showed that the variance of this ES in CRD is

$$\mathrm{Var}(d_W) = \left(\frac{k_T + k_C}{n k_T k_C}\right)\left(\frac{1 + (n-1)\rho}{1-\rho}\right) + \frac{d_W^2}{2k_T k_C (n-1)}, \tag{24}$$

where d_W is the ES statistic using the individual-level standard deviation. Even though this ES has a noncentral *t* distribution, it can be assumed to be approximately normally distributed (Hedges, 2007b). The formula of the standard error of d_W when the numbers of individuals in each group are not equal was also provided by Hedges (2007b).

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. Cheung (2009) also showed that likelihood-based CI's have advantages over Wald-CI's. Unfortunately, programs that compute likelihood-based CI's, such as Mx, do not support convenient methods of conducting multilevel modeling. Therefore, I will use the Wald CI by Mplus (Muthen & Muthen, 2007) to analyze the Wald CI of ES in multilevel analysis.

Sample Size Estimation by Power and Accuracy in Parameter Estimation in Two-Group CRD

As described above, in CRD, the sample sizes involved are cluster size and number of clusters. Therefore, different combinations of cluster size and number of clusters provide equal power or width of CI of ES. Therefore, researchers can pick many combinations. Most of the time, the combination of cluster size and number of clusters that result in the least cost is the preferred condition. To begin with, I will illustrate the cost function. Then, I will show how to find the optimal combination of sample sizes based on power and width of CI of ES.

Cost Function

When the sample size variables are only number of clusters and cluster size, the cost function can be expressed as

$$C = k \times \left(\text{cluster cost} + (n \times \text{individual cost})\right), \tag{25}$$

where *k* is number of clusters, and *n* is cluster size. Sometimes, the costs in treatment and control conditions are different. Therefore, the numbers of treatment and control clusters are not necessary to be equal. Therefore, the total cost changes as the sum of total treatment condition cost and control condition cost, which is

$$C = C_T + C_C, \tag{26}$$

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

$$C_T = k_T \times (\text{treatment cluster cost} + (n \times \text{treatment individual cost})) \tag{27}$$

$$C_C = k_C \times \left(\text{control cluster cost} + (n \times \text{control individual cost})\right). \tag{28}$$

In this case, when the costs of treatment and control conditions are different, the optimal cost from different combinations providing equal power or width is one which collects

more data from the less expensive condition. Large discrepancies between the numbers of clusters from each condition, however, require more stringent adherence to assumptions.

Sometimes, researchers may try to reduce the number of total individuals collected, instead of the cost. However, the number of total individuals collected is a special case of cost function when cluster costs are 0 and individual costs are 1.

Power Analysis

CRD power analysis is an extension of the independent t-test and ANOVA, which have different measures of ES. I will focus on the two-condition comparison only. The multiple condition comparison power analysis is shown in Barcikowski (1981).

As shown in the standard error equation above, the null hypothesis distribution is a central *t* distribution where

$$t = \frac{\hat{\gamma}_1}{\sqrt{\mathrm{Var}(\hat{\gamma}_1)}} \tag{29}$$

with $n - k$ degrees of freedom, where is $\hat{\gamma}_1$ the raw score mean difference. When the ES with the individual-level standard deviation is specified, the alternative hypothesis is distributed as a noncentral *t* distribution, as shown above.

Assuming that the individual level standard deviation is 1, the raw score mean difference ($\hat{\gamma}_1$) is equal to the ES estimate (d_W). Also, assuming that both null and alternative hypothesis distributions are normal, as shown in the Appendix, the variance of ES which provides a given power is

$$\text{Optimal Var}(d_W) = \left(\frac{d_W}{z_{1-\alpha/2} - z_{1-\text{power}}}\right)^2, \tag{30}$$

where d_{W} is ES estimate using individual-level standard deviation, and z_x represents quantiles in the normal distribution. Researchers can estimate the combination of cluster size, number of treatment clusters, and number of control clusters by solving

$$\text{Optimal Var}(d_W) = \frac{\sigma^2_{(Y|Z)} + n\tau_{(Y|Z)}}{nkp(1-p)}, \tag{31}$$

which is created by constraining Equation 30 and either Equation 12 or 21to be equal. Notice that an increase in the number of clusters decreases the variance of ES. The variance can approach zero. The cluster size, however, has a subtle relationship with the variance. The variance can be reduced by increasing cluster size to a certain degree, but its limit is not 0 (Rotondi & Donner, 2009).

The optimal cost combination of sample sizes may be found mathematically. Raudenbush (1997) showed that, when the number of treatment and control groups are equal, the optimal cluster size for the CRD model without a covariate is

$$n(\text{optimal}) = \frac{\sigma_Y^2}{\tau_Y} \cdot \sqrt{\frac{\text{cluster cost}}{\text{individual cost}}}, \tag{32}$$

where τ_Y and σ_Y^2 are cluster- and individual-level error variances, respectively. In a covariate CRD model, however, the optimal cluster size depends on the intraclass correlation of the dependent variable and the covariate, as well as the covariate effects on both levels (see Raudenbush, 1997, for a detailed discussion of the optimal cluster size for CRD with different kinds of covariates.).

Liu (2003) provided the optimal ratio of treatment and control conditions when the cluster-level costs of both conditions are not equal, assuming that the individual level of both conditions is equal. The optimal ratio is

$$\frac{p}{1-p} = \sqrt{\frac{\text{control cluster cost}}{\text{treatment cluster cost}}}, \tag{33}$$

where p is the proportion of treatment clusters.

In addition to the standard parametric approach, the empirical Bayes approach (Rotondi & Donner, 2009) and *a priori* Monte Carlo simulation is also available for power analysis in CRD (Muthen & Muthen, 2002). Both approaches are based on specifying parameter values in the population (e.g., intraclass correlation or raw score mean difference) and making hypothetical sampling distributions. The empirical Bayes approach can specify the intraclass correlation as a range of values. The empirical Bayes approach, however, is not directly linked to the individual-level ES. The advantage of the empirical Bayes approach is not essential in sample size estimation because researchers can set multiple intraclass correlations while searching for optimal values of sample sizes. Thus, I will explain only the priori Monte Carlo simulation, which is used in my program.

A priori Monte Carlo simulations build large numbers of samples from specified parameters. Next, the samples are used to calculate the desired statistics. Then, the easiest way to find a specified power is to check how many samples are statistically significantly different from the null hypothesis value. In this process, a large number of samples are required. Muthen and Muthen (2002) used 10,000 hypothetical samples for their analyses.

Many programs are available to calculate power in the two-group CRD, such as Optimal Design (Spybrook, Raudenbush, Congdon, & Martinez, 2009), PINT (Snijders & Bosker, 1993), ML-DE (Cools, Van den Noortgate, & Onghena, 2008), PASS (Hintze,

2008), R (Rotondi & Donner, 2009), or Mplus (Muthen & Muthen, 2002). Optimal Design, PINT, PASS and ML-DE are programs designed for finding sample size, but R and Mplus are generic packages that require researchers to write their own code, which can involve a steep learning curve. A disadvantage of PINT, PASS and ML-DE is that they require the researcher to specify raw score differences instead of standardized mean differences. Optimal design is the only program that accounts for standardized mean differences. However, it uses the total standard deviation, instead of the individual-level standard deviation. PASS can analyze the power of all regression coefficients for multilevel models; the cluster size, however, must be specified in advance, and PASS cannot find the optimal sample sizes by the cost function. PINT can analyze power of all regression coefficients also. Researchers need to specify a range of sample sizes, and PINT will calculate standard error for all combinations of the specified sample size range. PINT uses the cost function, as shown in Equation 25, to estimate an optimal cluster size given the number of clusters and a limited budget. PINT, however, cannot separate treatment and control condition costs. ML-DE also uses the cost function but, like PINT, cannot separate treatment and control condition costs. Furthermore, ML-DE requires a calculation process with several complicated steps involving multiple programs. Thus, PAWS that accompanies this thesis is designed to address the shortcomings of the currently available programs. PAWS finds a combination of sample sizes that has the largest power when researchers have a limited budget. In addition, PAWS is designed to be user-friendly and free.

Accuracy in Parameter Estimation

Researchers may need to know the minimum sample size to ensure that the ES is accurate enough. ES accuracy is measured by the width of CI of ES. As shown above, there are two strategies to analyze the CI of ES in two-group CRD. The analytical approach by Hedges (2007b) cannot be used directly in the models with covariates. Therefore, the phantom variable approach is preferred. To use the structural equation modeling package, however, the parameters from the model, such as ES, covariate effect, or intraclass correlation, cannot be put in the model directly. Fortunately, the parameters may be used indirectly by simulating Monte Carlo samples. Kelley (2008) used this approach to find the accuracy in parameter estimation of the coefficient of determination in multiple regression analysis. This approach is similar to the power analysis approach proposed by Muthen and Muthen (2002). First, the specified parameters are used to construct simulated samples. Next, the CI of ES is analyzed from all simulated samples. Then, the results of the width of CI of ES from all samples are used to find the average width. The degree of certainty (e.g., 80% of samples have the width of CI of ES less than .50) can be estimated by finding the percentile from the width of CI of ES data from all samples.

Before using the *a priori* Monte Carlo simulation approach, however, the sample sizes are required to make the samples. Therefore, the initial values of sample sizes are required. I will use the normal distribution approximation to find the initial values of sample sizes. As shown in the Appendix, the desired variance for a given width of the CI of ES is

$$\text{Optimal Var}(d_W) = \left(\frac{\omega}{2z_{1-\alpha/2}}\right)^2. \tag{34}$$

The series of combinations of sample sizes that provide the same optimal variance can be solved. Then, we can find the optimal cost combination and use it in *a priori* Monte Carlo simulation. *A priori* Monte Carlo simulation is used to adjust the combination of sample sizes to provide the more accurate results.

CHAPTER II

PROGRAM DESIGN

The main purpose of PAWS is to estimate the optimal sample sizes required in the two-group CRD using three criteria: the least number of individuals, the optimal cost given power (or width of CI of ES), and the maximum power (or the least width of CI of ES) from a fixed budget. Although PAWS can introduce a covariate in the model, one must assume that the covariate and the treatment variable are not correlated. In other words, there is no difference in the covariate means in two treatment conditions. Two types of covariates are available: individual-level and cluster-level covariates. In addition, researchers can estimate the sample sizes in which the numbers of treatment and control clusters are not equal. This feature is appropriate when the treatment and control conditions' costs are not equal, which compensates with loss of robustness when the parametric assumptions are violated. PAWS also allows researchers to fix some characteristics of sample sizes in advance: number of clusters, cluster size, or proportion of treatment groups. For example, researchers may know in advance that the classroom size is approximately 25. Therefore, researchers would like to find the number of clusters and proportion of treatment clusters only, given the costs of data collection. PAWS may be used to analyze the post hoc analysis, find power or CI of ES, given sample sizes, ES, and intraclass correlation.

The general approach of PAWS uses the normal approximation as initial estimates and *a priori* Monte Carlo simulation to provide more accurate results. When researchers would like to find the combination that provides the least expensive or the least number of individuals given power or width of CI of ES, several steps are involved. First, PAWS uses normal approximation to find many combinations of sample sizes that provides a given power or CI of ES. Then, PAWS picks a sample size combination that provides the least number of individuals or the least cost. Next, all parameters are used to make simulated samples (e.g., 10,000 samples) in different sample size combinations. PAWS makes a band of sample size combinations around the initial estimates of sample sizes. After the simulated samples of all sample size combinations around the initial estimates are built, PAWS checks whether the initial estimates provide accurate power or CI of ES with the least number of individuals or the least cost. If yes, PAWS will return the result. If not, it will adjust the band of sample sizes and reanalyze again until PAWS has found the sample sizes combination that provides accurate power or CI of ES with the least cost or the least numbers of individuals. In the fixed budget criterion, PAWS will find the various combinations of sample sizes and searches for the least variance of ES (maximum power and least CI of ES). Then, *a priori* Monte Carlo simulation is used to give more accurate results.

PAWS is designed to be user-friendly. PAWS is written for PC using Visual Basic 2008 Express. *A priori* Monte Carlo simulation samples are made by Mplus (Muthen & Muthen, 2007) with multilevel add-on (or combination add-on). Therefore, PAWS requires Mplus to estimate sample sizes by *a priori* Monte Carlo Simulation. Even without the Mplus multilevel add-on, PAWS still provides researchers with the

initial estimates of sample sizes. The degree of certainty, however, cannot be given because it requires the distribution of widths of CI of ES from simulated samples.

CHAPTER III

PROGRAM VALIDATION

I used PAWS 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. It requires many steps, however, 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. Some of the other programs are designed for only special cases, such as Optimal Design or ML-DE, or R code provided by Rotondi and Donner (2009).

I validated PAWS 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 by finding the sample sizes given

each situation. PAWS provided the power, width of 95% CI of ES, and width of 99% CI of ES. 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 30 and 34, respectively. I also used PAWS to find the power, width of 95% CI of ES, and width of 99% CI of ES, based on normal approximation, which is used to calculate starting value.

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 calculates the starting values accurately. The methods provided by Snijders and Bosker (1993) and the normal approximation, however, do not account for sampling error of intraclass correlation in the dependent variable. Also, these methods treat the covariate as a fixed effect. From now on, I will refer to the normal approximation method and the Snijders and Bosker method as the approximate method.

Accuracy of the Approximate Method

The power, width of 95% CI and width of 99% CI for the approximate 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.

In sum, statistical power estimated by the approximate method and the *a priori* Monte Carlo simulation is similar. The widths of CI of ES estimated by both methods are mostly similar. When the model includes a covariate with a low intraclass correlation, however, the dependent variable with high intraclass correlation and less than total sample size of 500, the width of CI of ES estimated by the approximate method is lower than the width of CI of ES estimated by *a priori* Monte Carlo simulation method. More systematic explorations are needed, however, to pinpoint exactly which situations produce large discrepancies between the two methods.

Table 1

Difference between the Approximate Method and the 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			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	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
All 5 Conditions	-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.

CHAPTER IV

CONCLUSION

The aim of this study is to develop an easy-to-use program to estimate sample sizes in CRD, based on either power or width of CI of ES. PAWS can estimate three characteristics of sample sizes simultaneously: number of clusters, cluster size, and proportion of treatment clusters. PAWS can estimate sample size combinations that provide the lowest budget given a specified level of statistical power or width of CI of ES. Also, PAWS can identify sample size combinations that yield the highest level of power, given a limited budget. Unlike other programs with similar functions, PAWS calculates an effect size that is standardized by the individual-level standard deviation. Thus, PAWS is particularly useful for researchers who wish to estimate the expected effect on an intervention on individuals in groups rather than the expected effect in terms of within-treatment group variance. PAWS uses two steps in sample size estimation. First, PAWS will find the starting values of sample sizes by using the normal approximation method. After that, the starting values will be used in the *a priori* Monte Carlo simulation to create simulated samples and determine whether the sample sizes combination has the desired characteristic. If not, PAWS will adjust the sample sizes combination until it achieved the desired characteristic. The *a priori* Monte Carlo simulation is better than the normal approximation method because the *a priori* Monte

Carlo simulation treats the intraclass correlation of the dependent variable as a random variable rather than as a fixed number. Also, the covariate in the model is also treated as random variable rather than as a fixed effect. Indeed, considering the intraclass correlation and the covariate as random characteristics is probably more valid because it accounts for the fact that these characteristics vary from sample to sample and can have dramatic and unexpected effects on power and CI of ES.

The results of PAWS were compared with PINT 2.2. The power and width of CI of ES calculated by the normal approximation method were the same as PINT 2.2 (within rounding error). The normal approximation method and the *a priori* Monte Carlo simulation method provide similar power and width of CI of ES. However, in some situations described above, the methods differ significantly in their estimates of the width of CI of ES. In such cases, the *a priori* Monte Carlo simulation method results are preferred.

Further development of PAWS may include options to consider the ES that is standardized by cluster-level or total-level standard deviation. PAWS may include how to account for multiple covariates. Also, the algorithm in PAWS may be improved to increase speed in estimating sample sizes by *a priori* Monte Carlo simulation, especially for estimating sample sizes combination based on degree of certainty of width of CI of ES. Finally, PAWS may be developed for other experimental designs, such as the multisite experiment (Raudenbush & Liu, 2000) or growth curve modeling.

REFERENCES

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational and Behavioral Statistics, 6*, 267-285.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13*, 99-109.

Bonett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods, 14*, 1-5.

Cheung, M. W.-L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling, 16*, 267-294.

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.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behavior Research Methods, 40*, 236-249.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.

Dalgaard, P. (2008). *Introductory statistics with R* (2nd ed.). New York: Springer.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

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.

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

Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*, 151-179.

Hedges, L. V. (2007b). 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.

Hintze, J. L. (2008). *PASS User's Guide II*. Kaysville, UT: Author.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in paramenter estimation via narrow confidence intervals. *Multivariate Behavioral Research, 43*, 524-555.

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.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110-129.

Liu, X. F. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics, 28*, 231-248.

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.

Muthen, L. K., & Muthen, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Muthen, L. K., & Muthen, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthen & Muthen.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

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

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*, 5-29.

Rotondi, M. A., & Donner, A. (2009). Sample size estimation in cluster randomized educational trials: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 34*, 229-237.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-259.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. Retrieved July 22, 2009, from http://sitemaker.umich.edu/group-based/files/od-manual-v200-20090722.pdf

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.

APPENDIX A

VARIANCE OF TREATMENT EFFECTS

First, I derive the formula of variance of treatment effects in the two-group CRD both with and without a covariate. In CRD, the dependent variable scores of each group can be partitioned in matrix notation as

$$\mathbf{Y_j} = \mathbf{X_j}\boldsymbol{\gamma} + \mathbf{1_n}\mathbf{u_j} + \mathbf{e_j}$$

where

$$\mathbf{Y_j} = [Y_{1j} \quad \cdots \quad Y_{nj}]^T$$

$$\mathbf{e_j} = [e_{1j} \quad \cdots \quad e_{nj}]^T$$

$$\mathbf{1_n} = [1 \quad \cdots \quad 1]^T$$

In the design without a covariate, the design matrix is

$$\mathbf{X_j} = \begin{bmatrix} 1 & X_j \\ 1 & X_j \\ \vdots & \vdots \\ 1 & X_j \end{bmatrix}$$

where X_j is 1 for treatment groups and 0 for control groups with

$$\boldsymbol{\gamma} = [\gamma_0 \quad \gamma_1]^T$$

$$\mathbf{u_j} = [u_{0j}]$$

In the design with a covariate, the covariate is partitioned into between- and within-cluster effects and the design matrix is

$$\mathbf{X_j} = \begin{bmatrix} 1 & X_j & Z_{ij} - \bar{Z}_j & \bar{Z}_j \\ 1 & X_j & Z_{ij} - \bar{Z}_j & \bar{Z}_j \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_j & Z_{ij} - \bar{Z}_j & \bar{Z}_j \end{bmatrix}$$

where X_j is 1 for treatment groups and 0 for control groups with

$$\boldsymbol{\gamma} = (\,\gamma_0 \quad \gamma_1 \quad \gamma_W \quad \gamma_B)^T$$

Assuming that the individual-level covariate effect (γ_W) does not vary across clusters, the model does not include the slope error. Therefore, the cluster error vector is

$$\mathbf{u_j} = [u_{0j}]$$

To simplify, the covariate has an overall mean of 0 and variance of 1. The variance of the predicted score is

$$\mathbf{Var}(\mathbf{Y_j}|\mathbf{X}) = \mathbf{V} = \tau_{(Y|Z)}\mathbf{1_n}\mathbf{1_n}^T + \sigma^2_{(Y|Z)}\mathbf{1_n}$$

Note that, in the model without a covariate, $\tau_{(Y|Z)}$ and $\sigma^2_{(Y|Z)}$ change to τ_Y and σ_Y, respectively. The matrix is

$$\mathbf{V} = \begin{bmatrix} \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} \end{bmatrix}$$

Matrix $\mathbf{V}$ has compound symmetry. That is, it has constant variance (all the elements of the diagonal are equal) and constant covariance (all the off-diagonal elements are equal). The inverse of a matrix with compound symmetry also has compound symmetry. Therefore, there are only two elements to find: on-diagonal and off-diagonal elements.

The first step is to find $\det(\mathbf{V})$. According to matrix lemma,

$$\det(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = (1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u})\det(\mathbf{A})$$

The **V** matrix can be partitioned as

$$\mathbf{V} = \sigma^2_{(Y|Z)}\boldsymbol{I} + \tau_{(Y|Z)}\mathbf{1}\mathbf{1}^T$$

$$\mathbf{A} = \sigma^2_{(Y|Z)}\boldsymbol{I}; \mathbf{u} = \tau_{(Y|Z)}\mathbf{1}; \mathbf{v} = \mathbf{1}$$

Thus,

$$\det(\mathbf{V}) = \left(1 + \mathbf{1}^T\left(\sigma^2_{(Y|Z)}\boldsymbol{I}\right)^{-\mathbf{1}}\left(\tau_{(Y|Z)}\mathbf{1}\right)\right)\det(\sigma^2_{(Y|Z)}\boldsymbol{I})$$

$$\det(\mathbf{V}) = \left(1 + \tau_{(Y|Z)}\mathbf{1}^T\left(\frac{1}{\sigma^2_{(Y|Z)}}\boldsymbol{I}\right)\mathbf{1}\right)\sigma^{2n}_{(Y|Z)}$$

$$\det(\mathbf{V}) = \left(1 + \frac{n\tau_{(Y|Z)}}{\sigma^2_{(Y|Z)}}\right)\sigma^{2n}_{(Y|Z)}$$

The on-diagonal elements of the inverse are

$$V_{11}^{-1} = \frac{(-1)^{1+1}}{\det(\mathbf{V})}\begin{vmatrix} \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} + \sigma^2_{(Y|Z)} \end{vmatrix}_{(n-1)\times(n-1)}$$

The determinant of the last matrix can be derived by matrix lemma, similar to $\det(\mathbf{V})$ with the result as

$$\left(1 + \frac{(n-1)\tau_{(Y|Z)}}{\sigma^2_{(Y|Z)}}\right)\sigma^{2(n-1)}_{(Y|Z)}$$

Therefore, the on-diagonal elements are

$$V_{11}^{-1} = \frac{\left(1 + \frac{(n-1)\tau_{(Y|Z)}}{\sigma^2_{(Y|Z)}}\right)\sigma^{2(n-1)}_{(Y|Z)}}{\left(1 + \frac{n\tau_{(Y|Z)}}{\sigma^2_{(Y|Z)}}\right)\sigma^{2n}_{(Y|Z)}}$$

$$V_{11}^{-1} = \frac{\sigma^2_{(Y|Z)} + (n-1)\tau_{(Y|Z)}}{\left(\sigma^2_{(Y|Z)} + n\tau_{(Y|Z)}\right)\sigma^2_{(Y|Z)}}$$

The off-diagonal elements of the inverse are

$$V_{12}^{-1} = \frac{(-1)^{1+2}}{\det(\mathbf{V})} \begin{vmatrix} \tau_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} + \sigma_{(Y|Z)}^2 & \cdots & \tau_{(Y|Z)} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} + \sigma_{(Y|Z)}^2 \end{vmatrix}_{(n-1)\times(n-1)}$$

The determinant of the matrix is the same as the determinant of a similar matrix which can be transformed by linear operations. Therefore, I subtract the second to the last rows by the first row as

$$V_{12}^{-1} = -\frac{1}{\det(\mathbf{V})} \begin{vmatrix} \tau_{(Y|Z)} & \tau_{(Y|Z)} & \cdots & \tau_{(Y|Z)} \\ 0 & \sigma_{(Y|Z)}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{(Y|Z)}^2 \end{vmatrix}_{(n-1)\times(n-1)}$$

$$V_{12}^{-1} = -\frac{\tau_{(y|x)}}{\mathbf{det}(\mathbf{V})} \begin{vmatrix} 1 & 1 & \cdots & 1 \\ 0 & \sigma_{(Y|Z)}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{(Y|Z)}^2 \end{vmatrix}_{(n-1)\times(n-1)}$$

$$V_{12}^{-1} = -\frac{\tau_{(y|x)}}{\mathbf{det}(\mathbf{V})}(1)(-1)^{1+1} \begin{vmatrix} \sigma_{(Y|Z)}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{(Y|Z)}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{(Y|Z)}^2 \end{vmatrix}_{(n-2)\times(n-2)}$$

$$V_{12}^{-1} = -\frac{\tau_{(Y|Z)}\sigma_{(Y|Z)}^{2(n-2)}}{\left(1 + \frac{n\tau_{(Y|Z)}}{\sigma_{(Y|Z)}^2}\right)\sigma_{(Y|Z)}^{2n}}$$

$$V_{12}^{-1} = -\frac{\tau_{(Y|Z)}}{\sigma_{(Y|Z)}^2\left(\sigma_{(Y|Z)}^2 + n\tau_{(Y|Z)}\right)}$$

Let

$$F = -\frac{\tau_{(Y|Z)}}{\sigma_{(Y|Z)}^2\left(\sigma_{(Y|Z)}^2 + n\tau_{(Y|Z)}\right)}$$

$$D = \frac{1}{\sigma_{(Y|Z)}^2}$$

The inverse matrix, $\mathbf{V}^{-1}$, can be written as

$$\mathbf{V}^{-1} = \begin{bmatrix} D+F & F & \cdots & F \\ F & D+F & \cdots & F \\ \vdots & \vdots & \ddots & \vdots \\ F & F & \cdots & D+F \end{bmatrix}$$

The variance of regression coefficients can be calculated by

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{J} \mathbf{X}_{\mathbf{j}}^{T} \mathbf{V}^{-1} \mathbf{X}_{\mathbf{j}} \right)^{-1}$$

This equation can be simplified in the two condition treatment effects as

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{k_T} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j + \sum_{j=k_T+1}^{k} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j \right)^{-1}$$

where k_T is the number of treatment clusters and k is the number of clusters. The $\mathbf{X}_j$ in the treatment and control condition have the treatment variable values of 1 and 0, respectively.

In the model without a covariate, the expression of treatment condition is

$$\mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j = \begin{bmatrix} n(D+nF) & n(D+nF) \\ n(D+nF) & n(D+nF) \end{bmatrix}$$

$$\sum_{j=1}^{k_T} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j = \begin{bmatrix} k_T n(D+nF) & k_T n(D+nF) \\ k_T n(D+nF) & k_T n(D+nF) \end{bmatrix}$$

The expression of control condition is

$$\mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j = \begin{bmatrix} n(D+nF) & 0 \\ 0 & 0 \end{bmatrix}$$

$$\sum_{j=k_T+1}^{k} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j = \begin{bmatrix} k_C n(D+nF) & 0 \\ 0 & 0 \end{bmatrix}$$

Therefore, the variance of the regression coefficients is

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{k_T} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j + \sum_{j=k_T+1}^{k} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j\right)^{-1} = \begin{bmatrix} kn(D+nF) & k_T n(D+nF) \\ k_T n(D+nF) & k_T n(D+nF) \end{bmatrix}^{-1}$$

The variance of the regression coefficient of difference between two condition means can be found in the element (2, 2), which is

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{kn(D+nF)}{kn(D+nF)k_T n(D+nF) - k_T n(D+nF)k_T n(D+nF)}$$

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{1}{nkp(1-p)(D+nF)}$$

When substituting D and F in the equation, the variance of treatment effect is the same the Equation 12, that is

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

In the model with a covariate, the expression of the treatment condition is

$$\mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j = \begin{bmatrix} n(D+nF) & n(D+nF) & 0 & \bar{Z}_j n(D+nF) \\ n(D+nF) & n(D+nF) & 0 & \bar{Z}_j n(D+nF) \\ 0 & 0 & D\sum_{i=1}^{n}\left(Z_{ij} - \bar{Z}_j\right)^2 & 0 \\ n\bar{Z}_j(D+nF) & n\bar{Z}_j(D+nF) & 0 & \bar{Z}_j^{\,2} n(D+nF) \end{bmatrix}$$

$$\sum_{j=1}^{k_T} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j$$

$$= \begin{bmatrix} k_T n(D+nF) & k_T n(D+nF) & 0 & n(D+nF)\sum_{j=1}^{k_T} \bar{Z}_j \\ k_T n(D+nF) & k_T n(D+nF) & 0 & n(D+nF)\sum_{j=1}^{k_T} \bar{Z}_j \\ 0 & 0 & D\sum_{j=1}^{k_T}\sum_{i=1}^{n}\left(Z_{ij} - \bar{Z}_j\right)^2 & 0 \\ n(D+nF)\sum_{j=1}^{k_T} \bar{Z}_j & n(D+nF)\sum_{j=1}^{k_T} \bar{Z}_j & 0 & n(D+nF)\sum_{j=1}^{k_T} \bar{Z}_j^{\,2} \end{bmatrix}$$

The expression of the control condition is

$$\mathbf{X}_j^T\mathbf{V}^{-1}\mathbf{X}_j = \begin{bmatrix} n(D+nF) & 0 & 0 & \bar{Z}_J n(D+nF) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & D\sum_{i=1}^{n}\left(Z_{ij}-\bar{Z}_j\right)^2 & 0 \\ \bar{Z}_J n(D+nF) & 0 & 0 & \bar{Z}_J^{\,2} n(D+nF) \end{bmatrix}$$

$$\sum_{j=k_T+1}^{k} \mathbf{X}_j^T\mathbf{V}^{-1}\mathbf{X}_j$$

$$= \begin{bmatrix} k_C n(D+nF) & 0 & 0 & n(D+nF)\sum_{j=k_T+1}^{k}\bar{Z}_J \\ 0 & 0 & 0 & 0 \\ 0 & 0 & D\sum_{j=k_T+1}^{k}\sum_{i=1}^{n}\left(Z_{ij}-\bar{Z}_j\right)^2 & 0 \\ n(D+nF)\sum_{j=k_T+1}^{k}\bar{Z}_J & 0 & 0 & n(D+nF)\sum_{j=k_T+1}^{k}\bar{Z}_J^{\,2} \end{bmatrix}$$

Therefore,

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{k_T}\mathbf{X}_j^T\mathbf{V}^{-1}\mathbf{X}_j + \sum_{j=k_T+1}^{k}\mathbf{X}_j^T\mathbf{V}^{-1}\mathbf{X}_j\right)^{-1}$$

$$= \begin{bmatrix} kn(D+nF) & k_T n(D+nF) & 0 & n(D+nF)\sum_{j=1}^{k}\bar{Z}_J \\ k_T n(D+nF) & k_T n(D+nF) & 0 & n(D+nF)\sum_{j=1}^{k_T}\bar{Z}_J \\ 0 & 0 & D\sum_{j=1}^{J}\sum_{i=1}^{n}\left(Z_{ij}-\bar{Z}_j\right)^2 & 0 \\ n(D+nF)\sum_{j=1}^{k}\bar{Z}_J & n(D+nF)\sum_{j=1}^{k_T}\bar{Z}_J & 0 & n(D+nF)\sum_{j=1}^{k}\bar{Z}_J^{\,2} \end{bmatrix}^{-1}$$

Because the variance of the covariate is 0 and the number of individuals is equal across clusters, the covariate expression can be simplified as

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left(Z_{ij}-\bar{Z}_j\right)^2 = SS_{W_Z}$$

$$\sum_{j=1}^{k} \bar{Z}_J^{\,2} = SS_{B_Z}$$

$$\sum_{j=1}^{k} \bar{Z}_J = k\bar{Z}.. = 0$$

Also, in the model, I assumed that the treatment variable and the covariate are not correlated. Thus, the average of the covariate in treatment and control conditions are zero, as

$$\sum_{j=1}^{k_T} \bar{Z}_J = k_T \bar{Z}_E = 0$$

Thus, the variance of regression coefficients can be simplified as

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\gamma}}) &= \left(\sum_{j=1}^{k_T} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j + \sum_{j=k_T+1}^{k} \mathbf{X}_j^T \mathbf{V}^{-1} \mathbf{X}_j\right)^{-1} \\ &= \begin{bmatrix} kn(D+nF) & k_T n(D+nF) & 0 & 0 \\ k_T n(D+nF) & k_T n(D+nF) & 0 & 0 \\ 0 & 0 & D \cdot SS_{W_Z} & 0 \\ 0 & 0 & 0 & n(D+nF)SS_{B_Z} \end{bmatrix}^{-1}\end{aligned}$$

The variance of difference between two conditions is the element (2, 2) of this matrix, which can be simplified by

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{(-1)^{2+2}}{\det\left(\mathrm{Var}(\hat{\boldsymbol{\gamma}})\right)} C_{22}$$

where C_{22} is the element (2, 2) of the conjoint matrix and

$$\begin{aligned}\det\left(\mathrm{Var}(\hat{\boldsymbol{\gamma}})\right) = kn(D+nF) &\begin{vmatrix} k_T n(D+nF) & 0 & 0 \\ 0 & D \cdot SS_{W_Z} & 0 \\ 0 & 0 & n(D+nF)SS_{B_Z} \end{vmatrix} \\ - k_T n(D+nF) &\begin{vmatrix} k_T n(D+nF) & 0 & 0 \\ 0 & D \cdot SS_{W_Z} & 0 \\ 0 & 0 & n(D+nF)SS_{B_Z} \end{vmatrix}\end{aligned}$$

$$\det\big(\mathrm{Var}(\hat{\boldsymbol{\gamma}})\big) = kn(D+nF)\big(k_T n(D+nF)\big)\big(D\cdot SS_{W_Z}\big)\big(n(D+nF)SS_{B_Z}\big) - k_T n(D+nF)\big(k_T n(D+nF)\big)\big(D\cdot SS_{W_Z}\big)\big(n(D+nF)SS_{B_Z}\big)$$

and

$$C_{22} = \begin{vmatrix} kn(D+nF) & 0 & 0 \\ 0 & D\cdot SS_{W_Z} & 0 \\ 0 & 0 & n(D+nF)SS_{B_Z} \end{vmatrix}$$

$$C_{22} = \big(kn(D+nF)\big)\big(D\cdot SS_{W_Z}\big)\big(n(D+nF)SS_{B_Z}\big)$$

Thus,

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{\big(kn(D+nF)\big)\big(D\cdot SS_{W_Z}\big)\big(n(D+nF)SS_{B_Z}\big)}{(k-k_T)n(D+nF)\big(k_T n(D+nF)\big)\big(D\cdot SS_{W_Z}\big)\big(n(D+nF)SS_{B_Z}\big)}$$

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{1}{nkp(1-p)(D+nF)}$$

When substituting D and F in the equation, the variance of the treatment effect is the same as the Equation 21, that is

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{\sigma^2_{(Y|Z)} + n\tau_{(Y|Z)}}{nkp(1-p)}$$

APPENDIX B

RELATIONSHIP BETWEEN TREATMENT EFFECTS AND PROPORTION OF VARIANCE EXPLAINED

This section will show the link between the regression coefficient and the proportion of variance explained. The cluster- and individual-level error variances of the models with and without a covariate can be linked as

$$u_j(\text{nulled}) = \gamma_B \bar{Z}_j + u_j(\text{covariate})$$

$$e_j(\text{nulled}) = \gamma_W \left(Z_{ij} - \bar{Z}_j\right) + e_j(\text{covariate})$$

Because I assume that the covariate and the treatment variable are not correlated and the average of the covariate is 0, the cluster variance of the null model can be partitioned as

$$\tau = \text{Var}(\gamma_B \bar{Z}_j) + \tau_{(Y|Z)}$$

$$\tau = \gamma_B^2 \tau_Z + \tau_{(Y|Z)}$$

and the error variance of the null model can be partitioned as

$$\sigma_Y^2 = Var\left(\gamma_W \left(Z_{ij} - \bar{Z}_j\right)\right) + \sigma_{(Y|Z)}^2$$

$$\sigma_Y^2 = \gamma_W^2 \sigma_Z^2 + \sigma_{(Y|Z)}^2$$

Therefore,

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

$$\gamma_B^2 = R_B^2 \frac{\tau_Y}{\tau_Z}$$

and

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

$$\gamma_W^2 = R_W^2 \frac{\sigma_Y^2}{\sigma_Z^2}$$

which are the same as the Equation 19 and 20.

APPENDIX C

STARTING VALUE OF THE OPTIMAL VARIANCE OF

THE DIFFERENCE BETWEEN CONDITIONS

Another equation to derive is the starting value of the optimal variance of the difference between conditions given power or width of CI of ES. I will assume the individual-level standard deviation is 1. It makes the individual-level ES equal to the raw score mean difference. I also assume that the ES is normally distributed. To find the starting value for a given power, the critical value of the null distribution equals the value in the alternative distribution in which the area under the alternative distribution and above the critical value that has a probability equal to power. Thus, the null hypothesis equation is

$$z_{1-\alpha/2} = \frac{\text{Critical Value} - 0}{\sqrt{\text{Var}(d_W)}}$$

and the alternative equation is

$$z_{1-\text{power}} = \frac{\text{Critical Value} - d_W}{\sqrt{\text{Var}(d_W)}}$$

When isolating the critical value of both equations and solving for the variance of the individual-level ES, the result is the same as the Equation 30, that is

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

To find the starting value given width of CI of ES, the distributions on the left and right hand side of the obtained ES are created. The centers of both distributions are the lower and upper bound, respectively. The right critical value of the lower bound distribution and the left critical value of the upper bound distribution equal the individual-level ES. Assuming that both distribution are normally distributed, the equation of the lower bound is

$$z_{1-\alpha/2} = \frac{d_W - \text{Lower Bound}}{\sqrt{\text{Var}(d_W)}}$$

and the equation of the upper bound is

$$z_{\alpha/2} = \frac{d_W - \text{Upper Bound}}{\sqrt{\text{Var}(d_W)}}$$

When isolating the individual-level ES of both equations and solving for variance of individual-level ES, the result is the same as the Equation 34, that is

$$\text{Var}(d_W) = \left(\frac{\text{Upper Bound} - \text{Lower Bound}}{z_{1-\alpha/2} - z_{\alpha/2}}\right)^2 = \left(\frac{\omega}{2z_{1-\alpha/2}}\right)^2$$