University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Open Access Theses and Dissertations from the College of Education and Human Sciences

Education and Human Sciences, College of (CEHS)

4-1-2011

The Effects of Simplifying Assumptions in Power Analysis

Kevin A. Kupzyk
University of Nebraska-Lincoln, kkupzyk2@unlnotes.unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/cehsdiss

Part of the Educational Psychology Commons

Kupzyk, Kevin A., "The Effects of Simplifying Assumptions in Power Analysis" (2011). *Open Access Theses and Dissertations from the College of Education and Human Sciences.* Paper 106.
http://digitalcommons.unl.edu/cehsdiss/106

This Article is brought to you for free and open access by the Education and Human Sciences, College of (CEHS) at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Open Access Theses and Dissertations from the College of Education and Human Sciences by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

THE EFFECTS OF SIMPLIFYING ASSUMPTIONS IN POWER ANALYSIS

by

Kevin A. Kupzyk

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

April, 2011

THE EFFECTS OF SIMPLIFYING ASSUMPTIONS IN POWER ANALYSIS

Kevin A. Kupzyk, Ph.D.

University of Nebraska, 2011

Adviser: James A. Bovaird

In experimental research, planning studies that have sufficient probability of detecting important effects is critical. Carrying out an experiment with an inadequate sample size may result in the inability to observe the effect of interest, wasting the resources spent on an experiment. Collecting more participants than is necessary may unnecessarily put too many participants at risk and potentially detect an effect size that is not clinically meaningful. Therefore, obtaining the most accurate estimates of necessary sample size prior to applying for research funding and carrying out experiments is of utmost importance.

Educational studies often select whole classrooms for participation. When clusters of individuals are assigned to experimental conditions, the design is referred to as a cluster randomized clinical trial. An over-estimation or under-estimation of the number of clusters needed can have large budgetary or logistical consequences. Accurate sample size estimates, especially in large-scale studies, can help researchers carry out high-quality research and make valid conclusions.

Statistical power is the probability of rejecting a false null hypothesis so that a researcher can correctly conclude that an effect has been observed in a study. Three different methods of estimating power are examined in this study, including (a) formulaic power functions, (b) empirical simulation studies, and (c) constructing exemplary datasets. Formula-based methods make assumptions that may not be met in practice. This study assessed (a) the extent to which failure to account for practical data conditions such as attrition, unequal treatment group sizes, and unequal cluster sizes bias estimates of anticipated power and sample size; and (b) if these effects were moderated by the amount of variability that is attributable to between-cluster differences.

The empirical simulation study and exemplary dataset methods showed that attrition and unequal treatment group sizes have substantial effects on estimates of power and sample size. Unequal cluster sizes did not appreciably affect power. Higher levels of bias were found when less variability was attributable to between-cluster differences. Power analyses based on a formulaic approach that fail to account for attrition or treatment group imbalance may severely overestimate power and underestimate the sample size necessary to observe important effects.

ACKNOWLEDGEMENT

I would like to thank my advisor, Dr. Jim Bovaird, for his hard work and support throughout my graduate career. His instruction and mentoring from the time I entered my master's program at the University of Kansas to the completion of my doctoral program at the University of Nebraska, has guided and motivated me to learn and achieve more than I thought possible. I would also like to express my appreciation to the members of my doctoral committee, Drs. Ralph De Ayala, Charles Ansorge, and Walter Stroup, for their feedback and help in guiding me through the process of completing my degree.

To my friends and family, thank you for your continued support and encouragement me along the way. This achievement would not have been possible without my family members, who have instilled in me the importance of education and cheered me on. I express my sincerest appreciation to my wife, Sara, who has been my greatest source of motivation, love, and support.

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION……..... 1

CHAPTER II. LITERATURE REVIEW……...……. 7

The Null and Alternative Hypotheses…….. 8

Type I Error……. 9

Type II Error……...……. 11

Power……...…...…. 12

Primary Factors Affecting Power……..... 13

Sample Size……..……... 14

Effect Size…… 15

Sources of Effect Size Evidence……. 17

Pilot Data…….……... 18

Meta-Analysis……..…….. 18

Interplay between Alpha, N, Standardized Effect Size, and Power…….…. 20

A Priori vs. Post Hoc Power……... 23

Unequal Variances and Sample Sizes…… 24

Effects of Complex Sampling…… 25

Analytic Models……...…….. 26

Units of Analysis…… 29

Intraclass Correlation Coefficient…….. 30

Degrees of Freedom……... 31

Financial Constraints……..... 32

Longitudinal Research – Repeated Measures…… 33

Attrition…….. 37

Methods for Determining Power……..……... 38

Subjects to Variables Ratio……..…… 39

Power to detect model fit…….…… 39

Power of the Likelihood Ratio Test….…… 40

Power Tables and Formulae…….……... 42

G*Power……..…… 45

Optimal Design…… 46

RMASS2……..…… 49

Exemplary Dataset Method……..……... 51

Empirical Power Analysis……..... 56

Number of Replications……..... 59

Current Study……..……. 64

Primary Research Questions……..…….. 66

CHAPTER III. METHODS AND PROCEDURES…… 68

Analytic Model…… 68

Choice of Power Analysis Methods…….……... 70

Simulation Conditions…….…… 72

Fixed Simulation Conditions…….. 72

Number of Occasions…….. 72

Effect Size……... 73

Manipulated Factors…….…... 74

Attrition……... 74

Treatment group balance……. 75

Cluster size ……. 77

Intraclass Correlation…….…….. 78

Summary of Experimental Design…….. 79

Procedures……... 80

Software for Empirical Power and Exemplary Dataset Power……... 81

Degrees of Freedom in Empirical and Exemplary Methods……... 82

Optimal Design……..…….. 83

CHAPTER IV. RESULTS…….……. 84

Initial Results……..……. 86

Research Questions 1……..……... 93

Research Questions 2…….. 95

Research Questions 3……..…... 95

Research Questions 4…….. 96

Increased Clusters Needed for Sufficient Power…….……... 97

CHAPTER V. DISCUSSION……. 99

Discussion of Research Questions……...…... 99

Degrees of Freedom …… 103

Comparison of the Empirical and Exemplary Dataset Methods…….…... 106

General Discussion…….. 110

References…….…... 113

Appendix A. SYNTAX FOR EMPIRICAL POWER ANALYSIS…… 124

Appendix B. SYNTAX FOR EXEMPLARY DATASET POWER ANALYSIS….. 127

Appendix C. SIMULATION DIAGNOSTICS……... 130

Appendix D. TYPE I ERROR CONDITION SUMMARY…… 133

LIST OF TABLES

Table 1. Exemplary dataset power estimates for each cell of the design…… 92

Table 2. Decreases in power as a result of attrition…….…..... 93

Table 3. Power estimates from adding clusters to account for attrition…….. 94

Table 4. Average declines in power due to violations at each level of ICC…..... 96

Table 5. Actual number of clusters needed for sufficient power……..……... 97

Table C.1-8. Simulation diagnostics…… 128

Table D.1-???Type I error condition summary……... 133

LIST OF FIGURES

Figure 1. Correct and incorrect decisions in null hypothesis testing……... 7

Figure 2. Example of using G*Power to calculate power for a t-test……..... 46

Figure 3. Example of using Optimal Design to calculate power for a t-test using HLM……. 47

Figure 4. Empirical power obtained from 500 replications across two seed values... 62

Figure 5. Experimental design…….…… 79

Figure 6. Power estimates under the .05 ICC condition…….……. 87

Figure 7. Power estimates under the .10 ICC condition.…….…… 88

Figure 8. Power estimates under the .15 ICC condition…….. 88

Figure 9. Exemplary power estimates under the .05 ICC condition…… 90

Figure 10. Exemplary power estimates under the .10 ICC condition…….. 91

Figure 11. Exemplary power estimates under the .15 ICC condition.……. 92

CHAPTER I. INTRODUCTION

In any experimental research setting, planning a study that will have a sufficient probability of detecting true effects is critical. The effects of carrying out an experiment that has a sample size that is too low or too high are well known. An inadequate sample size will likely lead to the inability to claim that the effect of interest has been reliably observed. Thus resources spent on an experiment may be wasted. In addition to wasting resources, over-sampling (i.e. collecting more data than is necessary) also has severe consequences, such as the potential to unnecessarily put too many participants at risk, withholding effective treatments from participants in control groups, putting unnecessary burden on participants, and potentially detecting an effect size that is too small to be clinically meaningful. Therefore, it is of utmost importance that the most accurate estimates of the necessary sample size be determined prior to applying for research funding and carrying out experiments. In fact, several major grant agencies, such as the National Institutes of Health, now require power analyses. When whole clusters are selected and assigned to experimental conditions, the design is referred to as a cluster randomized clinical trial. In educational studies involving students and classrooms, whole classrooms (i.e. *clusters* of students) are often selected for participation all at once. Any over-estimation or under-estimation could have large budgetary or logistical consequences. Accurate sample size estimates, in large-scale studies especially, can help researchers carry out high-quality research and make valid conclusions.

In order to accurately determine the necessary sample size, a researcher must carry out a power analysis. Statistical power is the probability of rejecting a false null hypothesis so that a researcher can correctly conclude that an effect has been correctly

observed in a study. Several different methods of estimating power are available, including (a) formulaic power functions, (b) simulation studies, (c) constructing exemplary datasets, (d) number-of-variables to number-of-subject ratios, and (e) likelihood ratio comparisons of nested models. Formula-based methods require assumptions to be made in practical data situations that may bias estimates of necessary sample size. When the assumptions of the power analysis will not be met in practice, the estimated value of power may be inaccurate, or biased. These assumptions, such as balanced sample sizes, homogeneity of variance, normally distributed data, independent observations or uncorrelated outcomes, and complete data are often unrealistic. In clustered data situations, where individuals are nested within clusters (e.g. students within classrooms), individuals may be more similar to others within their cluster than they are to participants in other clusters. Such clustering may invalidate the assumption of independent observations. The assumption of complete data is often violated in longitudinal studies. Planning a longitudinal study and not taking attrition into account in the power analysis may severely bias the actual sample size necessary to find an effect. Many researchers are confused by or simply unaware of the assumptions made and, subsequently, their research may be inadvertently compromised. Such a deficiency in understanding the effects of the realities of experimental studies on power may be impeding scientific progress in all areas of educational research and exacerbating budget problems that are currently pervasive in all research and development activities.

Every aspect of a planned experiment may have an effect on power, and consequently, on the necessary sample size estimation. Assumed effect size, source and amount of variability in scores, method of sampling, and the analytic design all have a

widely demonstrated effect on statistical power. Power analyses that do not adequately reflect the reality of the experiment and analyses to be conducted may provide a biased estimate of necessary sample size. The limiting assumptions made by power analysis software packages may have the effect of undermining the efforts of well-intentioned, but ill-informed researchers.

The purpose of this study is to investigate the circumstances in which methods of conducting power analyses will provide equivalent results as well as where, and to what extent, the methods diverge in their estimation of power. The *primary goal* of the study is to assess why, when, and to what extent estimates of optimal sample size are biased. Specifically, the formulaic, simulation-based, and exemplary dataset methods will be directly compared. The literature review will explore these and other common methods of power analysis, particularly focusing on their assumptions and limitations. Ignoring violations of research design assumptions will often have a deleterious effect on power, but the conditions that have the most noted effect and the degree of bias imposed are not well-documented. The study will attach numerical and logistical meaning to the penalties incurred by ignoring violations of common assumptions in power analysis, by examining change in statistical power and the number of participants.

This study will systematically examine the robustness of three available power analysis methods to common violations of the assumptions made. This can be carefully examined via extensive empirical power analyses. The computational burden and complexity associated with a well-conducted empirical power analysis is often prohibitive to applied researchers, leading to many limiting assumptions being accepted in their own power analyses. The primary desired outcome of this study is to examine the

sources and magnitude of biases inherent in typical approaches to study planning and power analysis. Specifically, this study will answer the following questions:

1. To what degree does failure to account for *attrition* bias estimates of sample size and power?
2. To what degree does failure to account for *unequal treatment group sizes* bias estimates of sample size and power?
3. To what degree does failure to account for *unequal cluster sizes* bias estimates of sample size and power?
4. Does the amount of *variability attributable to clusters (i.e. intraclass correlation, or* ICC) affect the magnitude of bias imposed by failing to account for these conditions?

The main effects of each of these conditions will be investigated as well as in combination. In the event that violation of the assumption does not have an effect on power, use of the formulaic power method would be appropriate. It is hypothesized that each of these three violations will have some effect on statistical power, but there will not be any moderating effects. It is also hypothesized that there will not be a sizeable difference in estimated power values obtained between the simulation-based and exemplary dataset methods of power analysis. The findings will indicate which of the assumption violations considered leads to the most severe loss in statistical power and provide direct estimates of the statistical power lost when these assumptions are violated.

Once the effects of violations on statistical power have been recorded, sample size, in each condition, will be increased using the simulation-based and exemplary dataset methods to determine how many more participants would be needed to achieve

sufficient power. This will provide a concrete idea of how many more participants are needed under each design violation and make the results more directly interpretable. Upon examination of bias in power caused by violations of assumptions, a critical discussion of the merit of each method of power analysis will be presented. The discussion will provide a significant contribution to the relationship between, and the understanding of, the different methods of power analysis typically used in research planning in the behavioral sciences.

A power analysis that makes the aforementioned assumptions when they will be violated in practice will lead to bias in the necessary sample size determination. Researchers planning atypical designs or those who wish to incorporate the realities they foresee in their experiment into their power analyses would thus benefit greatly from knowledge of which method best allows them to determine the necessary sample size and make sample size allocation decisions without injecting such biases into their study planning. Although many researchers would certainly like to make more accurate statements about power and necessary sample size, methods for exemplary datasets and empirical power analyses have not been utilized on a broad basis due to the intricacies of the methods and the fact that they have not been well investigated.

The context of interest is longitudinal randomized clinical trials, which are common in the field of education. Therefore, the current study does not address power for statistical tests in less complex experimental designs, such as t-tests, correlations, and ANOVAs in non-clustered, cross-sectional designs. It is assumed that traditional methods of assessing power for statistical tests in less complex experimental designs, such as t-

tests, correlations, and ANOVAs are quite accurate as there are fewer assumptions in these cases that may be violated in practice.

A final goal of the current research is that it will motivate readers to learn and use accurate, rigorous, and sophisticated methodologies for assessing statistical power within the randomized clinical trial context and experimental research in general. This will in turn improve the quality of future research proposals and improve the rigor of planning experimental research.

CHAPTER II. LITERATURE REVIEW

Power analysis is directly related to hypothesis testing in general. In a hypothesis test, we specify two hypotheses, a null hypothesis and an alternative hypothesis. Only one of the two hypotheses can be true in the population, thus, the hypotheses are mutually exclusive. Furthermore, because one of these two hypotheses must be true, they are exhaustive. As it is most commonly used in the social and behavioral sciences, the null hypothesis (H_0) holds that there is no effect in the population (e.g. the correlation or mean difference is zero). An alternative hypothesis (H_A) states that the H_0 is false; that there is a nonzero effect in the population. The alternative hypothesis is that which is consistent with the researcher's theory that a nonzero effect truly exists in the population. However, the researcher's theory is considered false until it is demonstrated that the null hypothesis is false beyond reasonable doubt (Kraemer, 1985). If the null hypothesis is, in fact, false, and it is statistically rejected, then a correct decision has been made. Statistical power is the probability of rejecting the null hypothesis when the null hypothesis is false. Likewise, if H_0 is true and it is not rejected, then a correct decision is also made. Figure 1 shows the two ways in which a correct decision can be made and the two ways in which an incorrect decision (error) can be made.

		Null Hypotheis	
		True	False
Decision	Reject H_0	Alpha (false rejection - Type I error)	Power (correct rejection)
	Fail to reject H_0	(correct retention)	Beta (incorrect retention - Type II error)

Figure 1. Correct and incorrect decisions in null hypothesis testing.

Each box in the figure has an associated probability value. The darkened boxes in the figure represent correct decisions. If the null hypothesis is true and the decision is made that there is no effect observed in a study (i.e. failed to reject H_0), the null hypothesis is correctly retained. If the null hypothesis is truly false, there is an effect in the population. If the effect is found to be significant, the null hypothesis has been correctly rejected. The probability of correctly rejecting a false null hypothesis is the *power* of the test. To gain an understanding of power and power analysis, one must be knowledgeable about the following concepts: the null and alternative hypotheses, Type I and Type II error, power, and the study design factors that influence the hypothesis testing system.

The Null and Alternative Hypotheses

The null hypothesis (H_0) states that there is no true effect in the population of interest. In a test of mean differences, it states that the involved groups have equal means. In a test of a relationship between two variables, it states that the correlation between the two variables is zero. The alternative hypothesis (H_A) states that there is a nonzero effect in the population. Under H_A, group means are assumed to be unequal, or a correlation between two variables is greater than or less than zero. The notion of a null hypothesis that should be accepted or rejected based on a statistical outcome was first proposed by Ronald A. Fisher in the early 1920s. Shortly thereafter, the concepts of an alternative hypothesis and two sources of error with nominal values (Type I and Type II errors) were put forth by Jerzy Neyman and Egon Pearson. For this reason, null hypothesis significance testing is often referred to as the Neyman-Pearson framework (Cowles, 2001).

Type I Error

Two types of errors are discussed in reference to the null hypothesis. If the null hypothesis is true and no effect is truly present, one hopes to have a very small chance of rejecting the null. The nominal error rate, alpha (α), is the maximum allowable probability of making a Type I error. Once alpha is determined, the critical value for the test statistic can be obtained. Alpha is typically set at $\alpha = .05$ *a priori* (prior to the study). The value of $\alpha = .05$ means we are willing to take a 5% chance ($\alpha*100$) of incorrectly determining that the null hypothesis is false. In other words, there is a .05 probability of incorrectly claiming there is an effect when there is not.

The nominal Type I error rate translates into the cut-off value for declaring statistical significance (i.e., $p < .05$). In this way, alpha is used to set up the region of rejection of a test statistic. The region of rejection is the range of values of a test statistic that indicate significant results. The critical value of the test statistic, which is the border value of the region of rejection, is the point at which the probability of the observed results (the p-value) is equal to the nominal value of allowable Type I error. In order to determine whether a test is significant, the p-value associated with the test statistic must be less than the value of alpha. The p-value is interpreted as the likelihood of the observed test statistic if the null hypothesis is true. In order to reject the null hypothesis, the p-value must be low, and alpha is the upper limit for the p-value deemed to be acceptable as an indication of significance. More stringent values of alpha (closer to zero) imply that it is less desirable to find effects to be significant when they do not actually exist, and therefore more extreme critical values must be reached in order to determine an effect to be significant. An example may be found in drug research. With experimental

drugs, the probability of determining a drug to be effective when it is not should be minimized, such that the likelihood of unnecessary side effects or health risks is reduced. Alternatively, less stringent values of alpha (higher than .05) may be acceptable if the focus is on increasing power and decreasing the likelihood of Type II errors. Determining that a potentially beneficial drug is not worthy of continued research and development may be of great concern. In the social sciences, however, less stringent values of alpha are not often implemented, as reviewers will not typically accept more liberal levels of Type I error (Hevey & McGee, 1998). An alpha level that is considered marginal (.05 to .10) may be seen as an indicator that an effect is trending towards significance and that further research is warranted.

Arguably, the nominal value of alpha is arbitrary. However, unless there is a strong justification for another value and full knowledge of what the consequences will be, alpha should always be set at .05 (Cohen, 1988). The $\alpha = .05$ level has been used since the 1600s with the advent of study into the normal distribution, although it has only been widely applied in hypothesis testing since the early 20th century, gaining popularity through the work of R.A. Fisher (Cowles & Davis, 1982). Although it is most often implemented simply out of tradition, one practical reason why .05 became popular is that in the normal, or Gaussian, distribution, it is the point at which a statistic has fallen roughly two standard deviations away from the mean. In terms of odds, $\alpha = .05$ means we are willing to accept a 1 in 20 chance of making a Type I error. It is the point above which the probability of making a Type I error was historically considered too high (e.g. .1), and below which the probability was considered too stringent.

Alpha is typically divided in half, with one half being placed at the lower end of the null distribution, and the other half at the upper end. Some of alpha is allocated to each end, or "tail," of the null distribution. Such "two-tailed tests" allow for effects to be positive *or* negative (i.e. bi-directional hypothesis). In a two-tailed test of mean differences with $\alpha = .05$, under the null hypothesis there is a .025 probability of finding one group's mean to be significantly *higher* than the other group's mean and a .025 probability of finding it to be *lower* than the other. If the investigator is sure the effect will be in the assumed direction and wishes to make no provision for the possibility of an effect in the opposite direction, a one-tailed test may be justified, and the entire amount of alpha is placed in the direction of the alternative hypothesis. Many have argued that most theoretical assertions are directional (Burke, 1953); for example, that children in a treatment group will demonstrate *higher* outcomes after an intervention than children in a control group. However, the possibility always exists that the effect may be in the opposite direction. It is also important to consider the acceptability of opting for a one-tailed test in a specific field of study. Such practice is often seen as a lowering of standards, and have gone so far as to say that one-tailed tests should only be performed when a difference in the opposite direction of that predicted would be psychologically meaningless (Kimmell, 1957). In general, two-tailed tests are the accepted standard.

Type II Error

If the null hypothesis is false, which is assumed in a power analysis, this means that an effect truly is present in the population. Determining that there is not a significant effect in observed data when there should be is a Type II error. Beta (β) is thus the acceptable probability of incorrectly determining that the null hypothesis is true. In other

words, the null hypothesis should have been rejected but it was not. The probability of a Type II error is not set explicitly. It is set indirectly by deciding on a desired level of power.

Power

The complement of Type II error is power (i.e. 1-β = Power). Since power and Type II error are complementary, an increase in power is, by definition, a decrease in β. Power is interpreted as the probability of correctly determining the null hypothesis to be false, or the probability of correctly rejecting the null hypothesis. Power has also been referred to as the sensitivity or true positive rate (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Determining that there is a significant effect in observed data when there should be is the meaning of power. It is important to note that statistical significance does not necessarily imply that an effect is clinically important. If an effect is important, however, statistical significance provides the empirical support needed to justify its scientific investigation.

One reason for performing a power analysis is to compare the statistical power of different experimental designs or analysis methods using a fixed sample size. Assuming the design and statistical test are decided upon in advance, which is most often the case in social science research, the primary reason for performing a power analysis is to determine how many participants are needed in order to have a high probability of rejecting the null hypothesis when it is false (i.e., declaring statistical significance). This probability is set in advance to a level that is typical in a certain field of study. In most social science research projects, a probability of .80 is the desired level of power. Experiments should have at least an 80% chance of finding an important effect to be

statistically significant, and thus a 20% chance of making a Type II error. The primary reason that this, albeit arbitrary, value became standard is the standard .05 probability for making a Type I error, which was traditionally seen as four times as serious as making a Type II error (Cohen, 1988, p. 56), thus Type I error probability is set at one fourth of the Type II error probability.

In some settings, higher levels of power may be necessary (e.g. 90%). When there is a high level of risk associated with falsely determining there is no effect present, one may want to minimize the probability of making a Type II error (i.e. $\beta < .20$). In the field of epidemiology, for example, which studies the effects of programs and interventions on public health, experiments are typically required to be carried out with 90% power (Bird & Hall, 1986). However, 80% power is considered adequate in most settings (Cohen, 1988; Houle, Penzien, & Houle, 2005; Leidy & Weissfeld, 1991), especially in the social sciences, and is therefore the most widely accepted value at which to set power. If a social science researcher submits a proposal in which power is set at 90%, reviewers would likely feel that the power is excessive and that their resources would be wasted by obtaining more participants than are absolutely necessary.

Primary Factors Affecting Power

Without changing the experimental design, statistical power is a function of sample size, effect size, variability, and alpha. Alpha is nearly always set at .05.. In designs typically seen in social science research, effect size refers to an observed mean difference between two groups. The ratio of effect size to variability is referred to as *standardized effect size*. However, standardized effect size is most often labeled *effect size*. In some applications, raw score units are used to express effect sizes, when the scale

used represents an important and interpretable metric, such as response time (Robey, 2004). Some authors recommend that the smallest mean difference in raw score units that is deemed clinically or substantively meaningful be used as the effect size (Hevey & McGee, 1998; Man-Song-Hing et al., 2002). However, effect sizes in social science research are primarily based on widely varying metrics, and mean differences are most often standardized by the standard deviation of the specific measurement tool. Therefore, the two primary factors affecting power considered here are sample size and standardized effect size.

Sample Size. In non-clustered, cross-sectional designs, the number of observations or participants in a sample is what is known as the sample size. All other determinants of power being held constant, increasing the sample size will always result in an increase in power. Larger samples lead to lower variance of the sampling distribution (often referred to as sampling variability or the standard error) and thus a more precisely estimated statistic. Less sampling variability increases the chance of finding a relationship or mean difference to be significant. Sampling variability can be seen as the inverse of precision, or sensitivity (Murphy & Myors, 2004). Population parameters are more precisely estimated with larger samples. Thus, studies with larger samples have more sensitivity and are more able to detect effects. Smaller samples provide statistics that are less stable, have more sampling error, and are thus less sensitive to detecting non-zero effects. The standard error of a statistic reflects the standard deviation of the statistic under repeated samplings under a given sample size (sampling error). As sample size is increased the sampling distribution for the statistic becomes less variable, and the standard error of the statistic is decreased. Repeated samplings with

large samples will produce a more narrow sampling distribution than with small samples, which is why smaller mean differences can be detected with larger sample sizes.

Effect size. In order to determine power, a researcher must have an idea, or an assumption, as to the magnitude of the effect that is assumed to exist in the population of interest. The degree to which the null hypothesis is false is reflected in the effect size. Determining the assumed effect size is widely regarded as being the most difficult part of conducting a power analysis. Effect size is generally interpreted as a scale-free measure, or a standardized effect.

Experiments are often designed in which participants are randomly assigned to one of two conditions. After the experiment, we may wish to assess the difference in means of the two groups' outcome scores. In tests of mean differences, the standardized effect size is the raw mean difference divided by the standard deviation of the outcomes, such that the difference is standardized and comparable across variables with different metrics. An effect size value, in this format, is interpreted as a percentage of the standard deviation of the variable of interest. Cohen's *d* (Cohen, 1988) is the metric most commonly associated with this measure of effect size and is calculated using the following formula:

$$d = \frac{m_1 - m_2}{\sigma} \tag{2.1}$$

where σ is the within-population standard deviation, which is assumed to be equivalent across groups. A similar metric is Hedges' *g* (Hedges, 1981), which replaces the within-population standard deviation with a pooled standard deviation, using group-specific sample sizes and variances to standardized mean differences. In tests of correlations, Pearson's *r* is already standardized, or unit-free, so the correlation *r* is the effect size. In

tests of main effects and interactions in an analysis of variance, Cohen's *f* is one of the most common measures of effect size, reflecting the standard deviation of standardized population means (Cohen, 1988), and is calculated as follows:

$$f = \frac{\sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}}{\sigma} \tag{2.2}$$

where *m* is the combined population mean, and *k* is the number of groups or levels of a particular factor. The ratio reflected in *f* is the standard deviation in group means to the standard deviation of within-population scores, and it can take on only positive or zero values. In the two-group case ($k = 2$), *f* reduces to .5**d*.

Two other measures of effect size that may be used in ANOVA models are directly related to *f* (Cohen, 1988, p. 281). Squaring Equation 2.2 results in:

$$f^2 = \frac{\sigma_m^2}{\sigma^2} \tag{2.3}$$

The ratio of group-mean variance to within-population score variance represents the percentage of variance accounted for by *group* membership. The percentage of variance accounted for by *population* membership is reflected by η^2, which a direct transformation of f^2 using the following formula:

$$\eta^2 = \frac{f^2}{1 + f^2} \tag{2.4}$$

The two effect size measures just mentioned, f^2 and η^2, are used when there are more than two groups (i.e. control, treatment A, treatment B). Cohen (1988) provides equations for converting *f*, f^2, and η^2 to *d*, which aids in interpretation of the relative size of the observed effects. Multiple R^2 is often used to reflect the percentage of variance in the

outcome accounted for by the set of predictors in a model. If the group indicator is the only variable included in the ANOVA model, multiple R^2 is the percentage of variance accounted for by *population* membership (Cohen, 1992) and is equivalent to η^2. A related effect size measure, omega-squared (ω^2), is interpreted in the same way as η^2, but is intended to provide a less biased estimate under small sample sizes (Olejnik & Algina, 2003). Omega-squared is also intended to reflect the characteristics of the population as opposed to the obtained sample.

There are many more effect size measures available. However, the measures presented here are those most commonly used in psychological and educational research. The type of statistical test is important to the determination of effect size, as the index should be defined in units appropriate to the data in question. Cohen's *d* is the metric used most often for tests involving normally distributed outcomes and dichotomous predictors, such as treatment indicators. Pearson's *r* is used to describe a relationship between two continuous variables, and thus is not used to describe experimentally manipulated group effects.

Sources of Effect Size Evidence

The main sources for determining what the assumed effect size will be are pilot data and past research in the form of published studies or meta-analyses. Using pilot data or published literature, an important aspect to this determination is the closeness of the measurement tools to be used and those used in the previous research. The similarity between the target population and the population previously studied is another important consideration.

Pilot Data. Assuming the measures and population are similar, calculation of the observed effect sizes in previous studies will be specific to the analytic model utilized in the previous studies. If the pilot data available fits the design and planned analysis of the study to be completed, the planned statistical test is applied to the pilot data in order to estimate the effect size, or parameter estimate and variance components, that may be expected in the experiment to be completed. Say, for example, we intend to estimate the effect of an intervention targeting the quality of parent-child interactions in low-income families. Prior to conducting a large-scale study, a small, local sample of 20 low-income families is collected and each is randomly assigned one of two treatment conditions: intervention or control. After the intervention is applied to half of the families, all are measured using a scale created to reflect family interaction quality. The means and standard deviations of each group are then used to estimate the mean difference attributable to the intervention and the amount of variability in the scores. The effect size is then calculated and used in planning the large-scale study as the size of the effect that may be expected in the population of low-income families.

Meta-Analysis. In addition to pilot data, previously-conducted research projects similar to the one being planned are a good source of needed information. A meta-analysis is a study of the published literature to assess typical effect sizes in a specific area of interest. The average effect size among studies that pertain to a specific construct and/or population can be used as the expected effect size in a planned study. For example, if a study is being planned to assess the impact of an early reading intervention on a sample of preschool children, we can survey early-reading literature to find effect sizes among studies of similar interventions. A meta-analysis may actually be a more

accurate method of determining an effect size than using pilot data, because a single dataset is sample-specific. Sampling multiple effect sizes could provide a much more stable and accurate estimate of the effect size that will be found in the study to be performed.

Upon conducting a meta-analysis, effect sizes must be placed on the same metric before they can validly be combined. For example, some studies may report effect sizes in the metric of Cohen's *d*, while others report in the metrics of *r*, *f*, or η^2. All effect sizes should be converted to the same metric before pooling the results. Another consideration that should be made prior to combining effect size estimates is whether or not the estimates represent the same population parameter (Morris & Deshon, 2002). For example, effect sizes from single-group pretest-posttest designs should not be combined with effect sizes reflecting differences between independent groups.

Upon reading or conducting a meta-analysis it is important to note that such results will tend to overestimate the effect size in the population. Studies that find small effects are less likely to report the statistics necessary for effect size calculations and are less likely to be published than those with large, significant effects (Greenwald, 1975). Thus, the mean or median of a range of effect sizes found in scientific literature will tend to overestimate the true effect size. Hedges (1984) provided a method to correct effect size estimates for the effects of censoring non-significant results. When meta-analyses are conducted, we typically assume the values collected are a random sample from the population of all possible effect sizes. However, what actually results is a left-censored distribution (not sampling the lower effect size values).

In the absence of pilot data or prior research on the relationship of interest, a power analysis must rely on theory or, as a last resort, arbitrary conventions of small, medium, and large effects (Murphy & Myors, 2004). Cohen's *d* values of .2, .5, and .8 are considered small, medium, and large, respectively. Correlation coefficients of .1, .3, and .5 are considered small, medium, and large, respectively (for a complete list of effect size metric conventions, see Cohen, 1992, Table 1). Relying on these conventions, however, is "an operation fraught with many dangers (Cohen, 1988, p. 12)." The perceived importance of effect size values across different fields should be a consideration. For example, a standardized mean difference of .2 may be seen as very significant in some disciplines, such as drug research, while it is seen as of little practical importance in others. Therefore, it is important to have some understanding of the effect sizes typically seen in a particular field of study as well as an understanding of what is practically-important, or "clinically-significant." In order to assess the adequacy of Cohen's conventions for the social sciences, Cooper and Findley (1982) reviewed social psychology literature and found the highest concentration of effect sizes to be around d = .5. Although the small, medium, and large conventions may be seen as arbitrary, there is an empirical basis behind them, and a medium effect size will often be seen as a reasonable assumption in the absence of pilot data or a meta-analysis.

Interplay between Alpha, N, Standardized Effect Size, and Power

Alpha, sample size, standardized effect size, and power are the four determinants of a closed system of power. Within the power system, these determinants can interact with and compensate for each other. Holding two of the four determinants constant, changing a third determinant will have a predictable effect on the fourth element. Power

in a t-test for independent mean differences, is calculated using the following formula given by Cohen (1988, p. 544):

$$Power = 100 * cdf\left(\frac{d(n-1)\sqrt{2n}}{2(n-1)+1.21(z_{1-\alpha}-1.06)} - z_{1-\alpha}\right) \tag{2.5}$$

where *cdf* is the cumulative distribution function of the unit normal curve, *d* is the effect size, *n* is the per-group sample size, and $z_{1-\alpha}$ is the critical value of the unit normal curve under the null distribution. When sample size and alpha are held constant, increasing the effect size will always increase power because it is found only in the numerator. Consider an example in which the per-group sample size is $n = 60$, the standardized effect size is $d = .4$, $\alpha = .05$, and therefore the critical $z_{1-\alpha}$ value is 1.96 with a two-tailed test. Using these values, statistical power is 58.4%. Increasing *d* from .4 to .5 gives a power value of 77.4%. Similarly, if effect size and alpha are held constant, increasing the sample size will always increase power because it is found in a higher power, or degree, in the numerator than in the denominator. Continuing the example, if instead of increasing *d*, *n* is increased from 60 to 70 participants per group, power increases from 58.4% to 65.2%. In order to achieve the same power associated with increasing *d* from .4 to .5 with $n = 40$ (77.4%), the per-group sample size would need to be $n = 93$. Over 100 more participants would need to be collected to achieve the same increase in power associated with an increase in effect size from .4 to .5.

A third situation exists in which effect size and sample size are held constant, and alpha is increased, causing an increase in power, or decreased (towards zero), causing a decrease in power. Continuing with the example, suppose we are less willing to make a Type I error, so alpha is decreased from .05 to .01, keeping *n* constant at 60 and $d = .4$. The result is an increase in the critical value ($z_{1-\alpha}$), making it more difficult to reject the

null hypothesis, and therefore power decreases from 58.4% to 33.9%. Alpha does not often deviate from .05 in the social sciences. However, an unexpected way in which alpha may be modified is in deciding if the researcher will be conducting a one-tailed or a two-tailed test. If the test is one-tailed, the full amount of alpha is placed in one tail of the probability distribution, thus lowering the critical value of the test statistic and increasing power. With $\alpha = .05$ in one tail, as opposed to two, which produced a power value of 58.4%, power has increased to 70.3%.

Another possibility exists in which power and alpha are held constant, and sample size is adjusted to determine the effect size that is detectable with the fixed level of power (typically 80%). This is often referred to as a "sensitivity analysis" (Faul, Erdfelder, Buchner, & Lang, 2009). In this situation, increasing the sample size results in a decrease in the effect size that is detectable with the set level of power. In research settings, this situation occurs when we are limited to a certain fixed sample size for budgetary or availability reasons. In this case, it is possible to determine how large an effect size can be detected with sufficient power. The formula for a sensitivity analysis involves algebraic manipulation and use of the inverse cumulative distribution function. The formula used here is as follows:

$$d = \frac{\left(cdf^{-1}\left(\frac{Power}{100}\right) + z_{1-\alpha}\right)\left(2(n-1) + 1.21(z_{1-\alpha} - 1.06)\right)}{(n-1)\sqrt{2n}} \tag{2.6}$$

where d is now the effect size detectable with given values of power, alpha, and sample size. Following the earlier example, where $d = .40$, $n = 60$, and $\alpha = .05$ in a two-tailed t-test resulted in power of 58.4%. Using this formula we can determine the detectable

effect size at 80% power. With a fixed sample size of 60 participants per group, there is 80% power to detect an effect size of .516.

A Priori vs. Post Hoc Power

The difference between *a posteriori* (retrospective) and *a priori* (prospective) power lies not in the calculations, but in how they are used in the research process. A priori power analysis determines how many participants are needed in the planned experiment, given the assumed effect size and desired level of power. Post hoc power determines the amount of statistical power, given an observed effect size and sample size in a completed experiment. In other words, post hoc power determines the likelihood that the results of an experiment would have been found to be significant. A post hoc power analysis is performed only after an experiment has been completed. In contrast, a priori power analysis is the determination of the necessary sample size to achieve a set level of power, given an assumed effect size, and is appropriate before an experiment is carried out.

One may encounter studies that report post hoc power along with non-significant findings as an explanation of why the results are non-significant. For example, an author might say "the results are non-significant, but power was found to be .43 and therefore we only had a 43% chance of finding this effect to be significant." Although some might see post hoc power as an explanation for non-significant results, such practice should be strongly discouraged. A more likely explanation is that the effect size did not turn out to be as large as expected. In fact, a lower than expected effect size is exactly what causes post hoc power to be low, assuming the sample size was sufficient. Many researchers and methodologists (e.g. Lenth, 2001; Hoenig & Heisey, 2001) argue against the use of

posterior power analysis. Power statistics from post hoc analyses often mislead the audience and are used to embellish data that has been collected (Littell et al, 2006). An example of this is a study that reports post hoc power along with non-significant results as an explanation of why the results are non-significant. However, it is important to note that both the observed power and the p-value are dependent on the observed effect size. Observed power is always low when results are non-significant. Nakagawa and Foster (2004) refer to this as the "power approach paradox" and also discourage the use of post hoc power analysis.

Unequal Variances and Sample Sizes

While power estimates are generally robust to violations of equal variances (Cohen, 1988; p. 274), unequal sample sizes between experimental groups are known to decrease statistical power. Even with the total sample size held constant, unequal within-group sample sizes result in an increase in the pooled standard deviation used for assessing statistical significance. The increased precision with which the mean of the larger group is estimated does fully compensate for the decrease in accuracy associated with the mean of the smaller group. The *effective* per-group sample size is often referred to as the *harmonic mean* of sample sizes, and is calculated as follows:

$$n' = \frac{2 * n_t * n_c}{n_t + n_c} \tag{2.7}$$

where n_t and n_c are the treatment and control group sample sizes, respectively. When groups are unequally sized, the effective sample size decreases as the ratio of one group's size to the other diverges, in either direction, from 1. As an example, suppose $n_t = 100$ and $n_c = 200$. Here, the control group's sample size is twice that of the treatment group (i.e. 2 to 1 ratio). The harmonic mean sample size is computed as 2*100*200/(100+200)

= 133, which is the effective per-group sample size, indicating that the sample of 300 participants will actually provide statistical power equivalent to a sample size of 266 equally-divided participants.

Effects of Complex Sampling

Complex sampling refers to cases in which data are collected in clusters, such as families, peer groups, classrooms, or towns. There are many instances in the social sciences when individual random sampling is feasible and commonly conducted. Such studies are often conducted in laboratory settings, whereas educational experiments are typically conducted in field settings. Experimenters will likely find it much more practical to sample intact clusters than individuals. Nested, or hierarchical, sampling are other terms that have been used to describe complex sampling (McDonald, 1993). For example, a common situation in educational research is when random samples of schools are selected from a population of schools, after which a random sample of classes are taken from within those schools, following with a random sample of students from the selected classrooms. The primary effect of complex sampling is the possible dependence between individuals. It is reasonable to assume that students are more similar to other students in their classroom than to students in other classrooms or schools, due to shared experience with a teacher or similar demographic characteristics.

Complex sampling is often coupled with methods of randomization of treatment groups that deviate from individual random assignment (i.e. randomized clinical trial; RCT). In situations where interventions are applied to teachers, students within a classroom cannot be assigned to more than one experimental condition without the risk of treatment contamination. The entire *cluster* is randomly assigned a treatment condition,

making it a *cluster* randomized clinical trial (C-RCT). Raudenbush (1997) refers to this type of experiment as a "cluster randomized trial." Aside from simply being more convenient, randomizing clusters is often the only feasible option when interventions are directed towards teachers. For example, teachers may be assigned to an intervention that targets teaching style or mode of classroom instruction. In this example, the focus of the study, teacher effectiveness, is at the higher level (i.e. teacher level), but lower-level, student outcomes are required to evaluate teacher effectiveness.

In addition to the previously discussed, and often-cited, determinants of power, any form of clustering within a dataset may affect statistical power through its effect on the standard errors used for tests of significance (Maas & Hox, 2005). To the extent that any variability between scores can be attributed to clustering of participants, the effective number of independent analysis units is decreased, and therefore power will be reduced. A more serious effect of such clustering is that the general linear model assumes cases are independent, and as a result, residuals are independent. Violating this assumption may seriously bias model estimated standard errors, and therefore, the statistical tests of parameter estimates.

Analytic Models. The general linear *mixed* model is able to account for such clustering by estimating variance components, in addition to the single residual variance, that capture variation in outcomes (intercepts) and bivariate relationships (slopes) that can be attributed to clusters. Multilevel modeling (Snijders & Bosker, 1999), mixed modeling (Littell et al, 2006), random coefficients regression (Longford, 1993), and hierarchical linear modeling (Raudenbush & Bryk, 2002) are all terms used to describe this method. The term mixed modeling is the most broad term, as it indicates that a

mixture of fixed and random effects are included in the model. However, the term hierarchical linear model (HLM) better reflects the clustered nature of the data. Therefore, HLM will be used throughout to refer to this class of modeling approaches.

In ordinary least squares (OLS) regression, a model for group differences in individual outcomes (Y_i) is as follows:

$$Y_i = \beta_0 + \beta_1 GROUP_i + r_i \tag{2.8}$$

where β_0 refers to the intercept (or mean of the control group if the group variable is coded 0 and 1), β_1 represents the mean difference between the two groups, and r_i is the individual's difference between observed and predicted scores (i.e. how far the actual score fell from the regression line). The set of residuals is assumed to be normally distributed with a mean of zero and variance represented by σ^2, the residual variance, which reflects the average squared distance between an individual's observed and predicted scores.

The OLS regression model is limited in that it assumes that cases are independent. Specifically, it assumes that residuals are independent. However, when individuals are clustered, there may be dependence in the residuals, in that individuals within a cluster are more alike than individuals between clusters. Failure to account for such dependence between scores may cause improper estimation of the variability between scores, bias in standard errors, and potential bias in estimates of effect size and power (Moerbeek, Van Breukelen, Berger, & Ausems, 2003). HLM accounts for the dependence by estimating extra variance components that capture the amount of variability in scores that is attributable to cluster membership. After conditioning on cluster memberships, the individual residuals may be considered independent. HLM is often represented by a

separate set of equations for each level of the hierarchy. Each fixed effect at a lower level has its own equation at the next level. For example, the level 1 (individual-level) equation for the simplest HLM extension of the previous example is as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} GROUP + r_{ij} \tag{2.9}$$

A subscript is added because individuals (i) are now nested within clusters (j). The level 2 equations for each fixed effect are:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{0j} &= \gamma_{01} \end{aligned} \tag{2.10}$$

where the intercept (β_{00}) is now a function of the grand mean (γ_{00}) and the difference between the cluster mean and the grand mean (u_{0j}). Variability in cluster means are often represented by the variance component, τ. The equations for each coefficient at the higher levels are then reinserted into the lower level equation to obtain the combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} GROUP_j + u_{0j} + r_{ij} \tag{2.11}$$

The combined form will be used throughout.

The use of HLM is widespread, meaning researchers are, increasingly, correctly accounting for the nested structure of their data. However, it is equally important that such clustering be accounted for in the power analysis stage of research planning. Typical power analyses for t-tests use degrees of freedom that assume individual-level random sampling and assignment. If a power analysis was conducted that did not account for clustering of the data, and the sufficient sample size was obtained according to that analysis when the actual data are clustered, the researcher may be in danger of obtaining biased results because non-independence of cases was not taken into account during study planning.

In a two-level, individuals within clusters framework, there are two sample sizes that should be considered. The level-one sample size, n or N_1, is the number of individuals within each cluster. The level-two sample size, J or N_2, is the number of clusters involved in the study. The total sample size is the sum of all J level-one sample sizes (J*n if all clusters are of equal size). The term sample size is most often used to refer to the number of subjects, or individual participants, in a study. However, the number of units of assignment is a more just description. A unit of assignment is that which is assigned to treatment conditions (Mead, 1988). In nested data situations, where individuals are nested within clusters such as classrooms or families, the cluster is the unit that is often assigned to a treatment condition. Typically, individuals within such clusters cannot feasibly be assigned randomly to treatment conditions.

Units of Analysis. An important distinction to make is between units of assignment, units of measurement, units of generalization, and units of analysis. The unit of measurement coincides with the level at which the outcome of interest is observed (e.g. student test scores). The unit of generalization is the level at which inferences are to be made. The unit of analysis is the lowest level at which scores are assumed independent, and thus treated as independent units (Kenny, Kashy, & Bolger, 1998). In many cases the units are the same. In complex sampling situations, however, the units may be different. Data may only be available in aggregate form, which changes the unit of measurement but not the unit of generalization. Numerous situations are possible. Kenny (1996) reviewed the issues involved in determining the proper unit of analysis in a multilevel context, depending on units of measurement and independence of cases.

Intraclass Correlation Coefficient. To assess the effect of clustering, the intraclass correlation coefficient (ICC or ρ) is used to measure the proportion of variance in the outcome attributable to clustering. In the two-level cross-sectional case, where scores from a single time point are collected on individuals nested within clusters, variance in outcomes can be separated into between-cluster variation (u or τ) and within-cluster variation (σ^2, r, or e), which is residual variation that is not related to cluster membership. The ICC formula is:

$$\rho = \frac{\tau}{\tau + \sigma^2} \tag{2.12}$$

The nature of this formula bounds the ICC between 0 and 1. Each of the variances takes on a value that is larger than zero, which means that the denominator will always be larger than the numerator. If there is no variability attributable to clusters, the ICC will be zero and all participants within and across clusters can be considered independent analysis units. However, when variability increasingly occurs at the cluster level, the ICC increases, and clustering has a clear effect on what can be considered the total number of independent analysis units, or the *effective sample size*. The effective sample size is determined by the *design effect*, which is a function of the ICC, and reflects the factor by which standard errors are underestimated in a clustered sample, by assuming a completely independent sample (Maas & Hox, 2005). In other words, it is the factor by which the variance of an estimate is increased by using a clustered sample, relative to a simple random sample (Snijders & Bosker, 1999). Zelin and Stubbs (2005) refer to the design effect as the extent to which the statistical reliability is made worse using a clustered sample relative to a similar-sized simple random sample. The design effect is calculated as follows:

$$DE = 1 + (n-1) * \rho \tag{2.13}$$

where n is the average cluster size. The lowest value the effect can take is 1, when the ICC is equal to zero. If the ICC is greater than zero, then the design effect is greater than 1, leading the effective sample size to be less than the total actual sample size.

The effective sample size is then:

$$N_{effective} = \frac{N_{actual}}{DE} = \frac{Jn}{DE} \tag{2.14}$$

where J is the number of clusters and N_{actual}, is the total actual sample size across all clusters. Examples of power calculations that involve clustered data will be presented below with the introduction of power analysis software.

Degrees of Freedom. In analytic models that take into account complex sampling, there are several choices available for the denominator degrees of freedom (DDF) used in testing fixed effects. Several popular software packages for analyzing these models have different default methods for calculating the degrees of freedom. The differences are most apparent when the effect being tested is an interaction between two variables measured at different levels of the hierarchy. The HLM software program (Raudenbush, Bryk, & Congdon, 2004) uses the number of units at the highest level involved in the interaction minus the number of fixed effects at that level of the model as the degrees of freedom, basing the determination only on the number of highest-level units. Several DDF methods available in SAS weight the number of lower-level units more heavily. The DDF used affects power by decreasing the critical value of the test statistic that must be reached to determine an effect to be significant. Determining DDF based on the number of lower-level units results in higher power estimates than using the number of higher-level units.

The five DDF options available in SAS using the MIXED procedure for HLM are residual, containment, between-within, Satterthwaite, and Kenward-Roger. Brief descriptions of the methods were presented by Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009) and by Schaalje, McBride, and Fellingham (2001). Assuming a two-level model, the *residual* method subtracts the number of fixed effects from the number of level-1 units and is the default method when the model does not contain a random statement (only a residual variance is estimated). The residual method ignores the covariance structure of the models and is said to be appropriate only for large sample sizes. The *containment* method is the default when a random statement is used and is based on the frequency of appearance of the fixed effect in the random statements. Containment DDF should be used only when the experimental design is balanced. The *between-within* method provides an improvement by partitioning the degrees of freedom into between and within portions. *Satterthwaite* and *Kenward-Roger* adjustments are approximate, as opposed to exact, methods that are determined in part by the estimates of the variance components. The Kenward-Roger method is often used to adjust for small-sample approximations of the variance components. Approximate methods are considered to be more appropriate when unbalanced, complex sampling designs are used.

Financial Constraints

Studies often have fixed budgets available for collection of participants. In multilevel studies there may be a known cost associated with collecting each cluster and each participant within a cluster. While it may be the case that collecting additional individuals within a cluster is much less expensive than collecting more clusters of the same size, the variability in scores that is attributable to clusters plays an important role

in this determination. If there is variability due to clustering, keeping the total sample size constant, different configurations of the number of clusters and the number of individuals within a cluster will produce varying estimates of power. In what has been referred to as "optimal sample size" determination (Moerbeek, Van Breukelen, Berger, & Ausems, 2003) or "optimal design" determination (Raudenbush, 1997), given a known, fixed budget, the design effect or ICC, and known cluster and individual unit costs, we can determine the combination of cluster size and number of clusters that provides the highest power (the optimal allocation). Investigating the optimal allocation, however, assumes that the experimenter has control over the size of clusters. It may be the case that clusters have a fixed size that cannot be adjusted, such as a classroom or small group of individuals.

Longitudinal Research – Repeated Measures

Many experiments are conducted longitudinally so that the effectiveness of interventions may be assessed. Multiple measurement occasions are necessary in such circumstances such that statistical tests of effectiveness can account for baseline scores, or how the individuals scored prior to implementation of the intervention. Also, educational intervention research is often focused on affecting the rates at which children or students change. Such an aim requires a model that accounts for the dependence of repeated measurements within individuals and assesses the difference in slopes, or rates of change, between two experimental groups. In other words, researchers aim to assess a treatment effect on rates of change.

One measure that is commonly used in analyzing data from studies with two or more measurement occasions is repeated measures analysis of variance (RM-ANOVA).

By calculating group means at each time point, RM-ANOVA assesses group differences in average rates of change between the experimental groups. Several strict assumptions are made, however, that may not often be met in practice. Similar to ANOVA, RM-ANOVA cannot account for clustering above the individual level, because it assumes person-level residuals are independent. In dealing with repeated measures, RM-ANOVA assumes all measurement occasions are taken at equal intervals, meaning that the length of time that elapses between measurement occasions is equal across all time points. The same assumption precludes the accommodation of individually-varying length of time between measurement occasions. Unless participants are measured at equal intervals, which are the same across all individuals in a study, RM-ANOVA will produce biased estimates of change parameters. If these conditions are met and there is no clustering in the data above the individual level, RM-ANOVA is an acceptable method with which to analyze longitudinal data.

Similar to cross-sectional designs, where a hierarchical linear model can be implemented to account for multilevel data structures, HLM is also well-suited to analyzing data from longitudinal designs. The primary difference is that now the lowest level of the hierarchy is different time points, and the individual is at level 2. The need for a model that accounts for nesting is apparent, because scores from the same individual across time will obviously be correlated because they come from the same person. The mixed model of interest in longitudinal studies where individuals are sampled independently is as follows:

$$\begin{aligned} Y_{ij} &= \beta_{00} + \beta_{01} GROUP_j + \beta_{10}(Time_{ij}) + \beta_{11}(GROUP_j)(Time_{ij}) \\ &\quad + u_{0j} + u_{1j}(Time_{ij}) + r_{ij} \end{aligned} \tag{2.15}$$

which reflects a two-level model for change (*i* time points nested within *j* individuals) where each outcome, y_{ij}, is a function of both fixed and random effects. The fixed effects include the intercept (β_{00}), a group effect on the intercept or the initial mean difference between the groups (β_{01}), an effect of time for the control group or the control group developmental trend (β_{10}), and the difference in developmental trajectories between groups (β_{11}). The fixed effect for the group, β_{01}, will accurately reflect the group difference at baseline as long as time is centered to reflect the number of measurement occasions since baseline (i.e. 0, 1, 2, and 3 as opposed to 1, 2, 3, and 4). The parameter of interest in this model, the time*group interaction parameter β_{11}, may be described as the differential growth rate that exists within the treatment group over and above the developmental trajectory of the control group (Curran & Muthén, 1999). The random effects include a level-1 residual variance (r_{ij} or σ^2), and two level-2 variance components: the variability in intercepts across individuals (u_{0j} or $\tau_{\pi 0}$) and the variability in slopes across individuals (u_{1j} or $\tau_{\pi 1}$).

Effect size in longitudinal intervention studies utilizing HLM can be measured by the difference in linear change between the two experimental groups divided by the standard deviation of the slope values (Raudenbush & Liu, 2001). Effect size is thus calculated as:

$$d = \frac{\beta_{11}}{\sqrt{u_{1j}}}. \tag{2.16}$$

Through the use of HLM, this effect size measure accounts for the dependence between repeated measurements taken from the same individual. HLM estimates a slope value for

each individual and the standard deviation of the slope estimates is used to calculate effect size, rather than the pooled standard deviation of all available scores.

As mentioned previously, HLM is appropriate for modeling change because we can consider repeated measurements (level 1) to be nested within individuals (level 2). In longitudinal experiments that involve clustered data, the analytic model should be modified to reflect the three-level nature of the design. Now time points are nested within individuals, which are nested within clusters. No new fixed effects need to be included, but added variance components are needed to account for the variability in intercepts and slopes that may be attributable to clusters. The hierarchical linear model of interest, with the added components on the second line, is now:

$$\begin{aligned} Y_{tij} &= \gamma_{000} + \gamma_{001} GROUP_j + \gamma_{100}(Time_{tij}) + \gamma_{101}(GROUP_j)(Time_{tij}) \\ &+ r_{0ij} + r_{1ij}(Time_{tij}) + \varepsilon_{tij} + u_{00j} + u_{10j}(Time_{tij}) \end{aligned} \tag{2.17}$$

Note that the symbols used for the fixed effects have changed from β to γ, to reflect three-level model conventions (Raudenbush & Bryk, 2002). There is also an added subscript, t, which refers to time points that are nested within i individuals, which are nested within j clusters. Treatment group is a cluster-level variable in a C-RCT. The level-1 residual variance is now represented by ε_{tij}, and the two level-2 variance components are: the variability in intercepts across individuals (r_{0ij}) and the variability in slopes across individuals (r_{1ij} or $\tau_{\pi 1}$). The level-3 variance components are: the variability in intercepts across clusters (u_{00j}) and the variability in slopes across clusters (u_{10j} or $\tau_{\beta 1}$). Effect size from this model (Spybrook, Raudenbush, Congdon, & Martinez, 2009) may be calculated as:

$$d = \frac{\gamma_{101}}{\sqrt{\tau_{\pi 1} + \tau_{\beta 1}}}. \tag{2.18}$$

where γ_{101} is the fixed effect for the time*group interaction and $\tau_{\pi 1}$ and $\tau_{\beta 1}$ are the individual- and cluster-level variances in slopes, respectively. The added random effects are important determinants of power because they affect how the raw score regression coefficient is standardized in the calculation of effect size.

Attrition

Power analyses for longitudinal studies often do not account for the possibility of participant dropout. When attrition is anticipated, power may be calculated based on the final number of participants that is expected to remain in the study at the final measurement occasion. In this situation, more participants are measured at baseline than are necessary, such that there will be enough participants at the end of the study to have adequate power. Such a strategy results in an underestimate of power. Participants that do not complete a study but did complete one or more measurement occasion still provide usable data, assuming that cases are not deleted in a list-wise manner.

It is often assumed that complete data will be collected from all participants. Such an assumption is not practical in some longitudinal studies involving human subjects. Although studies that are carried out over very short time periods could possibly be completed without loss of participants, some attrition can still be expected. Individuals or families often move or may simply choose not to continue participating in a study, as they are free to do at any time. The effect of attrition on power will depend on both the rate of attrition (the percentage lost between measurements) and the number of measurement occasions. For example, if a study with four occasions loses 10% of participants between each time point, only 70% of the original sample will remain at the

final occasion. If the analytic model deletes cases in a listwise manner, all the data obtained from the 30% that did not complete the study is inadmissible. When data are deleted listwise, the analysis may be referred to as a "completers-only" analysis. When data are not deleted listwise, power still decreases due to less available data at later time points, although there is less of a decrease. Data that are available at earlier time points can contribute to the analysis. Using all available data are consistent with an "intent-to-treat" analysis where treatment-control data are analyzed as intended or as randomized, regardless of treatment integrity, service use, or participation in training sessions (Jo, 2002; Atkins, 2009). Regardless of the analytic model, as attrition rates rise, power decreases. If no provision is made for the possibility of participant dropout, the necessary sample size at the beginning of a study will be underestimated to some degree. The experiment may fail to detect truly non-zero effects due to failure to account for attrition in study planning.

In some longitudinal studies where some attrition is expected, different attrition rates across treatment conditions may be a possibility. Researchers may expect individuals in an intervention to have less (or in some cases more) attrition over the course of a study. In these instances, time-specific sample size will be different across groups, and there will be some effect on power.

Methods for Determining Power

Power tables and formulae are commonly used power analysis methods in educational research settings. There are many other methods available, although most are inappropriate or inadequate for use in planning most educational studies. This section presents some of the well-known methods along with some of their advantages and

limitations. Use of power tables is then discussed, followed by more advanced methods for determining power.

Subjects to Variables Ratio. Some researchers have suggested using a ratio of participants to variables as a guideline for determining sample size. Obtaining 10 subjects per measured variable has been utilized in the past as a rough guideline. However, there are several determinants of power, as previously discussed, that are not accounted for by this method, most notably effect size. Green (1991) reviewed several of these guidelines put forth in prior research, compared them with results from traditional power analysis methods, and determined that a rule of obtaining $50+8m$ participants, originally proposed by Harris (1975) with a rule of obtaining $50+m$ participants, where m is the number of predictors, provided fairly accurate approximations for studies with medium effect sizes and less than seven predictors. However, it was observed that the number of participants needed is not linearly related to the number of predictors. The necessary sample size per predictor is greater with fewer predictors than with many. Green found such rules to be inflexible, rarely congruent with traditional power analysis methods, and generally argued against their use. Although using this approximation would be better than conducting no power analysis whatsoever, it is clearly a misguided and inadequate approach (Wampold & Freund, 1985). The major flaw inherent in using this guideline is that effect size and statistical power are not at all considered.

Power to detect model fit. Structural equation modeling (SEM) is a framework that estimates latent, or unobserved, variables and the structural relationships between them based on a set of observed variables. A power analysis method presented by MacCallum, Browne, and Cai (2006; earlier by MacCallum, Browne, & Sugawara, 1996)

computes power for overall model fit in SEM. Power analyses are typically focused on finding a single parameter to be significantly different from zero. Model fit refers to how well the hypothesized structural equation model fits the data. In other words, whether the observed data confirms, or is consistent with, the hypothesized model. The root mean square error of estimation (RMSEA) is a fit statistic often used in structural equation modeling. The RMSEA is based on the degree of non-centrality reflected in the chi-square fit statistic. An RMSEA value of .05 is considered "good", so it is often used as the null value (H_0: $\varepsilon \leq .05$). The alternative value chosen represents the lack of fit expected in the population. Common alternative values are .08 for "fair" model fit, or .10 for "poor" fit. The difference between the null and alternative values may be considered the effect size in this method (MacCallum & Hong, 1997), because it represents the degree to which the null hypothesis is false. To implement the method, the user must supply the model degrees of freedom, the null (typically .05) and alternative (typically .08 or .10) RMSEA values, and either the desired power or the sample size. By constraining the sample size, the achieved power can be determined. By constraining the desired power, alternatively, one obtains the sample size necessary to have sufficient power to detect poor model fit ($\varepsilon = .10$), if the model does fit poorly in the population. MacCallum, Browne, and Sugawara (1996) present power tables for tests of close fit ($\varepsilon = .05$ vs. $\varepsilon = .08$) at varying levels of sample size and model degrees of freedom, although power and sample size can be computed for any two RMSEA values.

Power of the Likelihood Ratio Test. Power for structural equation models can, however, be assessed for specific parameters. Satorra and Saris (1985) developed a test for power to detect model misspecification through the use of the likelihood ratio test.

The chi-square test statistic from a structural equation model is a measure of model misspecification. The likelihood ratio test compares two chi-square values: one from a null model, where the parameter of interest is freely estimated, and another from a nested alternative, or reduced, model in which the parameter is fixed at zero. The likelihood ratio test assesses the significance of the improvement in fit by freely estimating the parameter. The difference in chi-square values is tested against a chi-square distribution with one degree of freedom. If the difference exceeds the critical value, at $\alpha = .05$, then estimating the parameter does significantly improve model fit.

In order to carry out a power analysis using this method, one must first compute the model-implied covariance or correlation matrix of a baseline model, which is hypothesized to be correct, and then do the same for a model with the path of interest set to zero (Loehlin, 2004; pp 70-73). In order to obtain the baseline covariance matrix, the correct structural model with the parameter of interest in place, is established with the population values of all model parameters specified. Path rules are then followed to construct the implied covariance matrix for all observed variables. The same process is carried out for the alternative model, which excludes the parameter of interest. The covariance matrices and sample size are then used as inputs into a structural equation modeling software package. The resulting chi-square fit statistic or change in chi-square is considered the chi-square approximation, as it reflects the degree to which the alternative model deviates from the null model. Power is obtained by calculating the value of the chi-square cumulative distribution function with one degree of freedom and the appropriate critical value (e.g. 3.84 at $\alpha = .05$).

The method of Satorra and Saris (1985) is associated with structural equation models and is not typically applied to many of the experimental situations involved in educational intervention research. Hierarchical linear models for longitudinal studies, however, are equivalent to latent growth models, which are structural equation models for change. This leads to the possibility that, instead of the Wald (z), or F-test for statistical significance of a parameter using HLM, one could determine the significance of the fixed effect of interest via a likelihood ratio test using chi-square values. Under the null hypothesis, the likelihood ratio test is asymptotically equivalent to the Wald test. The asymptotic nature of the test, however, leads to high sampling variability at low sample sizes. As sample size grows large the likelihood ratio test and the Wald test should nearly always provide equivalent results, but there is no evidence of a point at which sample size is large enough to rely on the likelihood ratio test (Gonzales & Griffin, 2001). Another distinct drawback of the likelihood ratio test methods for power calculations is that they cannot account for missing data (Theommes, MacKinnon, & Reiser, 2010).

Power Tables and Formulae. The methods for determining power that have been described to this point are not commonly used in educational research settings. Finding power and necessary sample size with power tables is likely the most common method. Values in power tables are a function of power formulae (these are based on calculation of the *non-centrality parameter*, which is described later in the chapter). To use basic power tables, one must first specify alpha and directionality. Cohen (1988) provides six power tables for each type of test considered, for alpha levels of .01, .05, and .10 for both one- and two-tailed tests. The tables can be used in a post hoc manner by locating the row for the sample size of the study and the column for the observed effect

size. The value found in the cell is the observed power of the performed study. To use the power table in an a priori manner, one can determine the necessary sample size by locating the expected effect size and moving down the column until the desired power level is found. Alternatively, we can determine, for a given sample size, what effect size the study will have sufficient power to find significant (sensitivity analysis).

Cohen's t-test tables (Cohen, 1988) assume equal within-group variances and group sizes. However, in the event of unequal sample sizes, calculating the harmonic mean of the two group sizes is recommended to provide a useful approximation to power (Cohen, 1988, p. 42). Power values given by the tables in this case will be underestimated, as the tables use degrees of freedom based on equal sizes (degrees of freedom are too small; Cohen, 1988, p. 42). In the event of unequal variances, if group sizes are approximately equal, Cohen advises using the root mean square of the two variances. The following formula (Cohen, 1988, formula 2.3.2) demonstrates the root mean square calculation:

$$\sigma' = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}} \tag{2.19}$$

Thus, the effect size is modified by using the square root of the average of the two group variances in the denominator of the effect size calculation, which should provide an adequate approximation to power.

A noncentral distribution is the distribution of a test statistic when the null hypothesis is false. The statistical power of a test is found by determining the area of the noncentral distribution that is outside the critical value of the test statistic under the null hypothesis (the central distribution), which defines the criterion for significance (Cohen & Nee, 1987). The noncentrality parameter, often denoted by δ or, more commonly, λ, is

a function of the effect size and the sample size and represents the degree to which the alternative distribution deviates from the central distribution. Under the central distribution, the noncentrality parameter is zero. When the noncentrality parameter is not zero, the distribution of the test statistic is necessarily noncentral.

Cohen's power tables, as well as three software programs that will be described below, utilize noncentrality parameters to calculate power. The formula for the non-centrality parameter varies as a function of the type of statistical test that is to be performed. For an independent groups t-test, the noncentrality parameter is calculated as follows:

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{n * n}{n + n}} \tag{2.20}$$

where μ_1 represents the mean of the first group, σ is the pooled standard deviation, and n is the per-group sample size. The first half of Equation 2.20 is equivalent to Cohen's d. To calculate power, the noncentrality parameter is taken as the mean of the alternative distribution. The total area of the alternative distribution that lies beyond the critical value(s) of the test statistic under the null distribution is the statistical power of the test.

A distinct advantage to using modern power software programs is that no interpolation is necessary between tabled values of effect size or sample size. Conducting a power analysis using printed power tables with an expected effect size of .45 would necessitate interpolation between the .40 and .50 effect size columns. Cohen (1988) provides approximations that are deemed adequate, which are a function of the desired effect size and the sample size necessary for an effect size of .10, regardless of the effect size metric. This approach may be inaccurate due to the nonlinear relationship between the effect size metric and power, as well as sample size and power. Software programs

allow the user to enter the exact effect size or sample size expected and will return exact power values.

Another advantage of most power programs over printed power tables is that most tables, as previously mentioned, assume equal treatment group sizes. This requires the calculation of a harmonic mean sample size. Some power analysis programs, such as G*Power and PROC POWER in SAS, allow the user to input each group's sample size independently. A ratio of the two group sizes may also be supplied by the user in the event the goal of the analysis is to compute the required sample size, given the desired power, as opposed to computing power, given the intended sample size.

G*Power. G*Power (Erdfelder, Faul, & Buchner, 1996) is a program that provides power for simple statistics such as t-tests, F-tests, and correlations. There are many other programs available that perform similar functions to G*Power, such as SAS PROC POWER and GLMPOWER (SAS Institute Inc., 2010), Power and Sample Size (Hintze, 2008), Power and Precision (Biostat, Inc., 2001), and nQuery Advisor (Elashoff, 2007). Many of the alternative power analysis programs listed also facilitate power computations for logistic regression, which is not available in G*Power. However, G*Power is freely available and performs power analyses for the most common statistical tests implemented in social science research. Figure 2 shows an example of using G*Power to calculate power for the scenario mentioned earlier. A two-tailed t-test is to be performed on a sample with 60 participants in each of two treatment groups. An expected effect size of $d = .4$ has a power value of 58.4%.

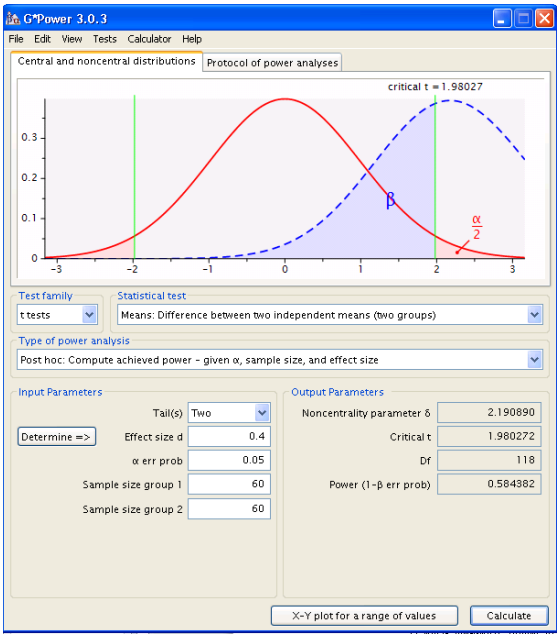

Figure 2 Example of using G*Power to calculate power for a t-test.

Power and sample size calculations may be performed for multiple regression and for many non-parametric tests, such as chi-square and Wilcoxon rank tests. For longitudinal studies, power for repeated measures ANOVA is provided. However, this program assumes equally spaced time points, no attrition, and completely independent cases (non-nested design). If there is any multilevel aspect to the data to be obtained, G*Power or any formulaic power analysis procedures that cannot account for clustering are not viable options for determining sample size requirements.

Optimal Design. Two formulaic power programs that are freely available compute power for longitudinal and multilevel analyses. Power in Two Level Designs (PINT; Snijders & Bosker, 1993) calculates the sample size necessary to achieve small enough standard errors for estimating coefficients with a desired degree of precision, not power. In educational evaluation studies, however, the focus is on detecting the

significance of regression coefficients of a certain magnitude. Optimal Design 2.0 (OD; Spybrook, Raudenbush, Congdon, & Martinez, 2009) provides power and sample size estimation in cross-sectional and longitudinal designs that assign treatments at either the individual- or cluster-level, with person- or group-level outcomes. The program allows the user to specify cluster size, alpha, the number of measurement occasions, the amount of variation accounted for by a cluster-level covariate, and the intra-class correlation coefficient, which is essential to assessing power in multilevel designs. Figure 3 shows power calculations for an extension the previous example that now assumes the 60 participants in each treatment condition are divided among 12 clusters with five individuals per cluster. Entire clusters are assigned to treatment conditions, making it a cluster randomized trial.

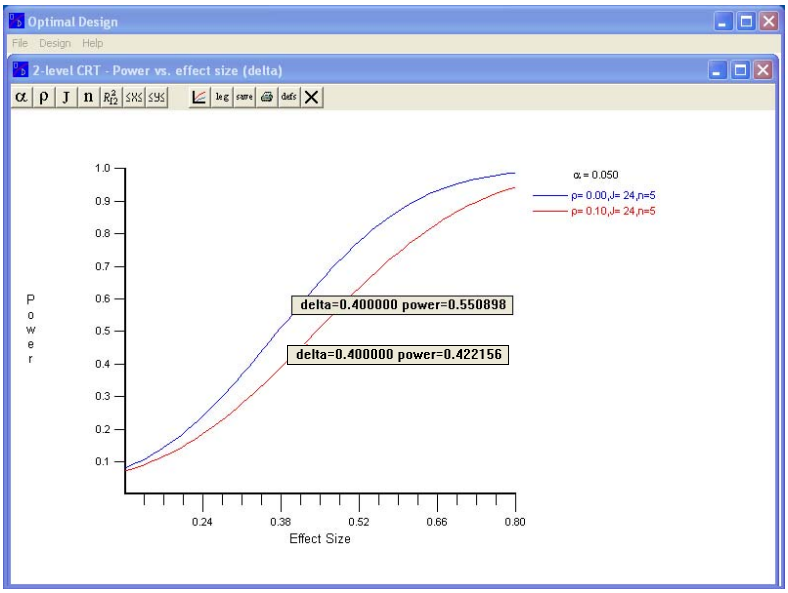

Figure 3 Example of using Optimal Design to calculate power for a t-test using HLM.

The two lines in Figure 3 represent two levels of the intraclass correlation coefficient: 0 and .10; that is, 0% and 10% of the variability in scores is attributable to clusters, respectively. When 0% of the variability is attributable to clusters, residuals may be assumed independent and the test of mean differences is at its highest possible level of power. As the figure shows, however, power to detect a standardized effect size of $d = .4$ is now 55.1%, reduced from 58.4%, found using G*Power. The standardized effect size is calculated as:

$$\delta = \frac{\bar{y}_E - \bar{y}_C}{\sqrt{\tau + \sigma^2}} \tag{2.21}$$

When the ICC is zero, τ is zero and Equation 2.21 reduces to the same formula used for Cohen's d (Equation 2.1). The slight reduction from the power value reported by G*Power is a result of sampling variation that leads to a slight, non-zero amount of variance at the cluster level. The lower line in the figure shows power values for the case where 10% of the variability in scores is attributable to clusters, which reduces power from 55.1% to 42.2%.

Although OD provides much greater flexibility in dealing with dependence among cases, the program assumes balanced groups, equal cluster sizes, homogeneity of variance, and no attrition. If the planned experiment will realistically meet these assumptions, then OD will provide an accurate estimation of the sample size necessary to achieve sufficient power.

Optimal Design determines power by using the design and variance components provided to calculate the noncentrality parameter (λ) for the statistical test of interest. The noncentrality parameter for a treatment effect is a function of the expected value of the parameter, the number of clusters, the cluster size, the number of time points, and the

intraclass correlation (a function of the residual variance and cluster-level variance). Under a two-level cross-sectional cluster randomized trial (i.e. a multilevel t-test), the F-statistic for testing treatment effects follows a non-central F distribution, F(1,*J*-2,λ), using the following formula

$$\lambda = \frac{J\delta^2}{4(\rho + (1-\rho)/n} \tag{2.22}$$

where *J* is the number of clusters, δ is the standardized effect size, ρ is the intraclass correlation coefficient, and *n* is the number of participants per cluster. Statistical power is then calculated by determining the proportion of the non-central F-distribution, F(1,*J*-2,λ), that falls outside of the critical value of the central F-distribution, F(1,*J*-2,0).

Under a longitudinal design, the power calculation must also take into account the number of measurement occasions. OD allows the user to set the frequency of measurements (*f*) and duration of the study (*D*). Frequency refers to the number of observations per unit of time, which is typically held at one (and will be throughout this study), while duration refers to the number of intervals that elapse in a study. The total number of time points, *M*, is equal to *f***D*+1. In a pre-post design, for example, there are two time points and therefore one time unit interval elapses in the study (i.e. *D* = 1). The duration is one and the total number of time points is two. Keeping all other things constant, adding measurement occasions increases power, because the sampling variability with which slopes are measured decreases (Raudenbush & Liu, 2001).

RMASS2. A distinct drawback to power programs that utilize power formulae is the assumption that sample size remains constant over the course of a longitudinal study. As mentioned previously, such an assumption is clearly unreasonable in any longitudinal study involving human subjects. One possible method to account for attrition is to use the

minimum number of expected cases at any (the last) time point and use a power formula or power software. Such a method will lead to an underestimate of statistical power when HLM or latent growth modeling is to be implemented. These analytic methods implement maximum likelihood estimation, which utilizes all available data from each respondent as opposed to deleting cases listwise.

Hedeker, Gibbons, and Waternaux (1999) addressed this issue by developing a program for sample size and power estimation in longitudinal studies that accounts for attrition as well as unbalanced experimental conditions and retention rates that vary between conditions. The program, called RMASS2, calculates sample size requirements for two-group longitudinal studies. The program was developed to compare linear trajectories of two groups over time, which is the goal of the time*group interaction in the HLM presented earlier. Variability in person-level intercepts and slopes is accommodated by the program. The RMASS2 program allows the user to input the number of time points, alpha, the number of tails to be used in significance testing, the desired level of power, the ratio of group sizes (i.e. accommodates unbalanced designs), the expected attrition rate, the number of random effects in the model and specify values for each of the effects in the case of complex variance-covariance matrix types, however the program has limited documentation and few available examples.

Another limitation of the RMASS2 program is that the retention rate is always assumed equal between the two experimental groups and is constant over the course of the study. In longitudinal studies with school-based samples, one may reasonably expect the rate of attrition to be higher between spring and fall (over the summer) than between fall and spring (Malmgren & Gagnon, 2005). The possibility also exists that participants

in the treatment group may drop out of a study at a lower rate than in the control group, or vice versa (Prinz & Miller, 1994). However, the primary drawback to the RMASS2 program is that no provision is made for nesting above the individual level. If the assumption of constant attrition is reasonable to a researcher and there will be no clustering of individuals in a planned study, the methods and RMASS2 program of Hedeker, Gibbons, and Waternaux (1999) may be of interest.

Exemplary Dataset Method. In order to instantly calculate power for an F-test, the non-centrality parameter must be known or calculable. The non-centrality parameter reflects the degree to which the alternative distribution deviates from the null distribution of F (the central distribution). The reason that calculating power from a formula is difficult in complex designs is that such designs render approximation of the non-centrality parameter exceedingly difficult, and thus, non-centrality formulas have not been derived for complex designs. A formula for the non-centrality parameter in every possible data collection situation is not possible, as there are innumerable ways in which studies can be designed.

O'Brien and Muller (1993) noted that the non-centrality parameter for a t-test is simply the *t* value that would be obtained if a researcher had "exemplary" data. An exemplary dataset is a set in which the mean of each group in the sample is precisely equal to the population mean of each of the groups and the standard deviation of the sample is precisely equal to the standard deviation in the population. To carry out a power analysis for a t-test, an exemplary dataset is created in which summary data, in the form of two means, two sample sizes, and one standard deviation, is submitted to an analysis program that will return the resulting F-statistic, such as SAS PROC GLM or

PROC GLIMMIX. The exemplary sums of squares hypothesis value, SSH, is needed to carry out the exemplary dataset method. Thus, the exemplary dataset in this situation is not actually a dataset, but a set of statistics describing the population. The F-statistic found after the t-test is performed is the non-centrality parameter, or the mean of the non-central F-distribution. The probability density function is then called to determine the proportion of the non-central distribution that lies beyond the critical value of the central F-distribution. The proportion obtained is the power of the test.

O'Brien and Muller provide SAS modules, originally titled OneWyPow (O'Brien & Muller, 1993), and expanded as UnifyPow (O'Brien, 1998), that are used for calculating power for a variety of statistical tests, which take as input the expected means, standard deviations, sample sizes, directionality, and alpha level. The modules return the expected power based on calculation of the NCP. For example, we can find the necessary components to calculate power for a one- or two-tailed t-test with 60 participants per group and an effect size of d = .4 using the following SAS syntax:

```
data;
input group $ Y n_exemp; datalines;
control   0 20
treatment 6 20
;
proc glm order=data;
class group;
freq n_exemp;
model Y = group;
run;
```

The dataset created has group labels, means, and sample sizes. The sample size given in this initial step is arbitrary, but the ratio between groups should reflect what will actually be observed. The source table printed by the GLM procedure contains the exemplary degrees of freedom and sums of squares necessary to run UnifyPow. Here, they are 1 and 360, respectively. The sums of squares value will change depending on the size of the

sample in the exemplary dataset. The following SAS syntax will complete the power analysis using UnifyPow:

```
%let UnifyPow = C:\Program Files\SAS\UnifyPow020817a.sas;
%include "&UnifyPow"; datalines;
exemplary SSH
NumParms 2
Nexemplary 40
sd 15
NTotal 120
effects
"control vs. treatment" 1 360
;;;;
%tables
```

In this syntax, 'NumParms' refers to the number of parameters in the GLM. There are two parameters in this example, the intercept and the group effect. The 'Nexemplary command specifies the total sample size of the exemplary dataset, 'sd' is the expected standard deviation of the scores, and 'NTotal' is the total planned sample size. The 'effects' statement specifies the degrees of freedom and sums of squares observed in the source table of the GLM results. Running the syntax shown produces a power value of .584, which is the same as that reported by the other methods in this example. O'Brien and Muller's methods assess power for general t, F, and chi-square tests, which are covered by G*Power and SAS PROC POWER.

For more complex general linear models, however, the modules do have applicability that is not covered by the other power programs mentioned. UnifyPow was created to handle models with random effects, as opposed to models with only fixed effects. The UnifyPow module can handle unbalanced sample sizes and directional tests. It is from this module that the term "exemplary dataset" method originates (O'Brien & Muller, 1993), however its use is different in this situation. For basic tests, such as t-tests, F-tests, and chi-square tests, which are covered by the software programs, "exemplary dataset" refers only to summary information. For general F-tests for linear models in

situations that fall outside of the more common statistical analyses covered by UnifyPow, a large dataset is simulated which characterizes and mimics the true state of affairs that is expected in the population, prior to initiating the module. In order for one to assume that the means and standard deviations are exactly equal to those expected in the population, the exemplary dataset must be very large, comprising hundreds of thousands of cases or more. A large sample size is needed to avoid variation in the non-centrality parameter due to its asymptotic nature (Saunders, Bishop, & Barrett 2003).

The next step is to submit the exemplary dataset to a linear model procedure, such as PROC GLM or PROC GENMOD, in order to obtain the exemplary sums of squares hypothesis value, SSH_e. The user provides this value along with the sample size of the exemplary dataset to UnifyPow. The process is explained by Saunders, Bishop, and Barrett (2003) in reference to case-control studies, a log-linear modeling situation. "By an exemplary dataset, we mean one in which the proportions of cases and controls in the different exposure categories take their expected values under the alternative hypothesis." The authors state that the method can be generalized for any alternative hypothesis as long as an exemplary dataset can be defined. Once the population parameters were determined, the authors simulated a large exemplary dataset (N = 200,000,000) and calculated the test statistic of interest. The test statistic is regarded as the non-centrality parameter under the alternative hypothesis with the large sample size. The parameter must then be scaled down to reflect the actual sample size that would be obtained in a real experimental situation. To calculate power for a given sample size the cumulative distribution function for the test statistic is called with the appropriate degrees of freedom and the rescaled non-centrality parameter obtained. However, Saunders, Bishop, and

Barrett (2003) only provide an example of a chi-square test, for which power can be calculated from several other programs.

Stroup (1999, 2002) demonstrated how SAS can be used to determine statistical power for a mixed model analysis to account for spatial variability (clustering). The procedure, although not described as the "exemplary dataset" method (O'Brien's previous work is cited by Stroup), involves constructing a dataset of the intended size (the anticipated number of participants) where each individual's score is a composite of only the fixed effects in the model. An important difference from the method described by Saunders, Bishop, and Barrett (2003) is that the method no longer involves any simulation of data. For example, if the mean of the control group is 0 and the mean of the treatment group is .3, then all individuals in the control group will have 0 and all in the treatment group will have .3 entered as their outcome value. The GLIMMIX procedure is then invoked to analyze the dataset, predicting the outcome with the group indicator, while holding the model variance parameters constant using the PARMS statement. The resulting F-value of the group indicator multiplied by the numerator degrees of freedom is the non-centrality parameter. The proportion of the central F distribution below the non-centrality parameter is the statistical power of the test. To demonstrate using the running example, the following SAS syntax creates a dataset with 60 participants in each group, where all entries in the control group have a value of 0 and all entries in the treatment group have a value of .4:

```
data; input y     group ne;datalines;
            0     0     60
            0.4   1     60
PROC GLIMMIX;
      CLASS group; freq ne;
      MODEL y=group/solution;
      PARMS (1)/hold=1;
      ods output tests3=fixed;
```

```
RUN;
data power;set fixed;
Fcrit=finv(.95,NumDF,DenDF,0);
ncp=NumDF*FValue;
power=100*(1-probf(Fcrit,NumDF,DenDF,ncp));
run;
```

The GLIMMIX procedure performs an F-test for the group effect, holding the variance to 1.0 with the PARMS statement. The F-value is then used as the non-centrality parameter to find the power, which is 58.4%, the same value found previously.

The method described by Saunders et al (2003) creates a much larger dataset with variability in the scores, then scales down the non-centrality parameter to what would have been obtained with a dataset of the intended size. While the two approaches are asymptotically equivalent, the Saunders et al. approach takes a very large number of cases to create stable estimates of variance components (often too large for a desktop computer's memory capacity). Stroup's method is more straightforward, in that there is no need to determine how to rescale the non-centrality parameter and it creates less of a burden on system resources.

Empirical Power Analysis

An approach that does not rely on a non-centrality parameter or cumulative distribution function is the empirical power analysis, also referred to as a simulation or Monte Carlo study (Littell et al, 2006; Muthén & Muthén, 2002). In this approach, a large number of datasets are generated to reflect random samples of a given size from the population of interest. For each sample drawn, the appropriate statistical test is applied and the significance or non-significance of the test is recorded. After all replications have been completed, the power value is obtained by calculating the proportion of the replications that had significant test statistics. If zeros were recorded for non-significant

tests and ones were recorded for significant tests, the mean of the zeros and ones give the proportion of significant replications.

How each dataset is constructed depends on the experimental design and the planned statistical test. However, a linear model equation is most common in situations of interest in the social sciences. For example, a dataset intended to be analyzed by an independent groups t-test can be generated by the following regression equation:

$$y = B(Group) + e \tag{2.23}$$

where B is the standardized regression coefficient, or effect size (d), e is the error term, and y is the generated outcome. Group takes on the values 0 and 1, so that the effect B is applied only to one group (i.e. the treatment group). The error term e is the most critical element of this approach. There must be a portion of the variance in y that is unrelated to the fixed effect of group. Otherwise, every data point within the same group will be equal. The error term represents the uncertainty involved when a sample is drawn from a population; the uncertainty is reflected by a statistical distribution (Murphy & Myors, 2004). After the group effect has been taken into account, variability is added in the form of a completely independent random draw from a normal distribution with a mean of zero and standard deviation equal to what is expected in the population of outcome scores (i.e. $N(0,\sigma^2)$). The random number generator in the computer program selects a random seed for each replication and is updated within each replication for each random draw. This ensures that no two simulated datasets will be the same. Therefore, the sampling error is what causes the sample to approximate a random sample from the distribution of interest.

In complex sampling situations where a hierarchical linear model is to be implemented, error will be separated into at least two components. To the extent that

some variability in outcomes can be attributed to cluster membership, there will be some non-zero value for between-cluster variance, the variance unexplained by treatment group membership that can be explained by cluster membership. The usual, lowest-level error term, *e*, is still in place to capture the remaining unexplained variability. The simplest equation for regression in this case becomes:

$$y = B(Group) + u + e \tag{2.24}$$

where *u* represents the between cluster variation in outcomes. The value of *u* is the estimated variance of the *J* cluster-means of the outcome variable. To obtain a specific intraclass correlation coefficient, *u* should be set proportional to *u* + *e*. The formula for the ICC (Equation 2.12) can be algebraically manipulated to easily determine the proper cluster-level variance for a given individual-level variance and desired ICC.

$$u = \frac{\rho * \sigma^2}{1 - \rho} \tag{2.25}$$

For example, if the variance of individual outcome scores is $\sigma^2 = 1$ and the expected ICC is $\rho = .1$, the cluster-level variance should be set at $u = .111$.

In longitudinal studies without clustering, the model from which data will be generated is:

$$\begin{aligned} Y &= \gamma_{00} + \gamma_{01} GROUP_i + \gamma_{10}(Time_{ti}) + \gamma_{11}(GROUP_i)(Time_{ti}) \\ &+ u_{0i} + u_{1i}(Time_{ti}) + r_{ti} \end{aligned} \tag{2.26}$$

where *r* is the time- and individual-specific variance, u_0 is the variance in intercepts across individuals, and u_1 is the variance in slopes across individuals. In a study with clustering, the resulting three-level model is used:

$$\begin{aligned} Y_{tij} &= \gamma_{000} + \gamma_{001} GROUP_j + \gamma_{100}(Time_{tij}) + \gamma_{101}(GROUP_j)(Time_{tij}) \\ &+ u_{00j} + u_{10j}(Time_{tij}) + r_{0ij} + r_{1ij}(Time_{tij}) + \varepsilon_{tij} \end{aligned} \tag{2.27}$$

where ε is now the time- and individual-specific variance, u_{00} and r_{0i} are the individual- and cluster-level intercept variances, and u_{10} and r_{1i} are the individual- and cluster-level slope variances.

Number of Replications. Muthén and Muthén (2002) provide a tutorial on carrying out an empirical power analysis, in which it is recommended that multiple analyses be performed using more than one seed value, in order to assess the stability of the estimates obtained. The reason this is important, although it was not acknowledged by the authors, is that the examples given in the study were carried out with a small number of replications, or simulated datasets.

In an empirical power analysis, one must determine the number of replications that will be completed. An apparent standard in published Monte Carlo studies is 1,000 replications (Koehler, Brown, & Haneuse, 2009). It is, however, logical to assume that the number of replications will have an effect on the precision in an estimate of power. In order to estimate the precision of a power estimate, it must be regarded as the mean of a binomial distribution. A binomial distribution is the sampling distribution of probability estimates that results from repeated samplings of sets of individual scores that can take on only one of two values (e.g. 0 or 1).

The accuracy of a binomial point estimate is defined in terms of its standard error. The consequence of this knowledge is that one can plan the number of replications in a power analysis so that the confidence interval for power estimates is sufficiently narrow, thus having the desired level of accuracy in the estimated power. Kelley (2007) used similar language in terms of accuracy in parameter estimation. However, the author was referring to planning the size of a sample to achieve a desired level of accuracy in a

sample statistic. Nonetheless, the number of replications in an empirical power analysis can be thought of in the same manner, as a sample size necessary to achieve a desired level of accuracy in a probability. The sample here, however, is the collection of *r* replications. The statistic is the estimate of power. The size of the sample, *n*, in each replication directly affects the magnitude of power (e.g. larger *n* leads to greater power), but *n* only indirectly affects the level of precision in the estimation of power. Holding *r* constant and increasing the sample size *n* will decrease the amount of sampling variation in the power estimate, assuming power is above .5, simply because the binomial distribution becomes less variable as the mean approaches 1. Decreasing *n* such that power drops from .5 to a lower value, also decreases the sampling variation in power. This is due to the fact that the standard error of a binomial distribution is at its peak when the mean is .5, as can be demonstrated using the formula for the standard error of a binomial:

$$s.e. = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{2.28}$$

where π is the probability of a certain response in the population and *n* is the sample size. This formula is only applicable when each of the *n* participants in a sample provide a Bernoulli trial response (i.e. 0 or 1). It terms of a power analysis, however, the sample size *n* only directly serves to increase or decrease the magnitude of power, or the mean of the distribution. For one replication, increasing *n* increases the likelihood of that replication (sample) receiving a 1 (i.e. the parameter of interest was significant) vs. a zero (non-significance of the parameter). The power is the average of these zeros and ones across all replications. It is then the number of replications *r* in the power analysis that has a direct effect on the precision of the power estimate, as each replication can be

thought of as a Bernoulli trial. The formula for the standard error of power may be expressed as:

$$s.e.(power) = \sqrt{\frac{\pi(1-\pi)}{r}} \tag{2.29}$$

Lachenbruch (2002) referred to the square of this equation as the observed variance of the power estimate, using p, where p is the observed power, and the number of replications in a simulation as the denominator. Yuan and Bentler (1999) used this formula to estimate the standard error of empirical Type I error rates. It is thus known that the number of replications defines the error in an empirical power estimate.

The standard error of power is the standard deviation of the sampling distribution of power estimates, given π and r. This means that if one wishes to plan a study to achieve 80% power, that is the value for the mean, π. Performing a standard number of replications of r = 1,000 would suggest the standard error for power is 0.0126, or 1.26%. A rough estimate for the 95% confidence interval may be obtained by $\pi \pm 2*s.e.$, which gives a confidence interval for power that is roughly 5% wide. The power estimate could fall at any point between 75.5% and 82.5%. It should be noted that since power is distributed as a binomial the true confidence interval is symmetric only when $\pi = .5$. Based on the binomial cumulative distribution function, the actual 2.5th and 97.5th percentile points, given $\pi = .8$ and $r = 1000$ are .775 and .824, respectively.

The following figures present results from an empirical power analysis for a linear regression of an outcome variable on a continuous predictor. In this example, the number of replications is 500, and n ranges from 71 to 81. The number of replications in this example was chosen to clearly elucidate the effect of a using a low number of replications.

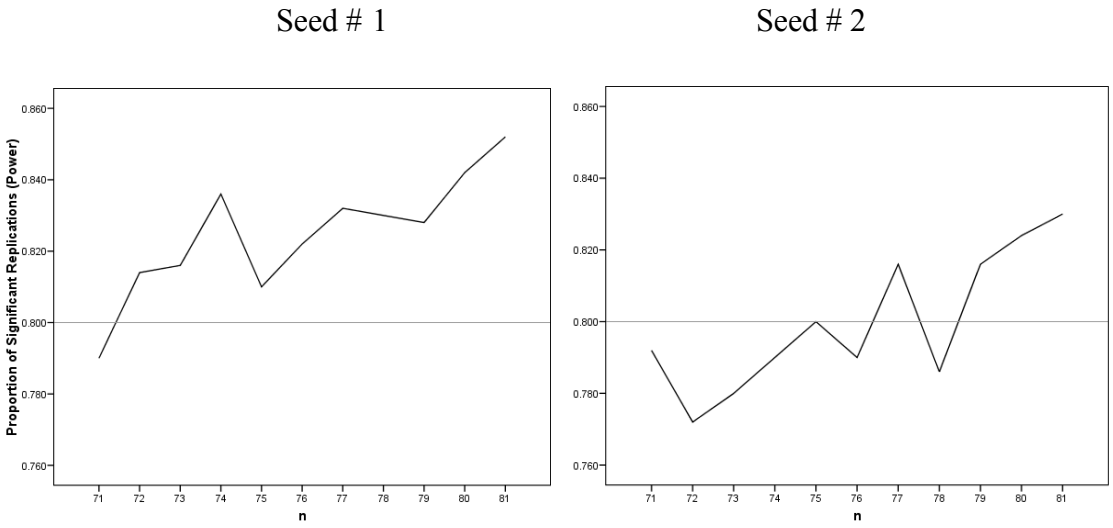

Figure 4. Empirical power obtained from 500 replications across two seed values.

The left panel shows the results from one randomly chosen seed value, and the right panel shows power results from a second seed. If one were to start with the first seed and increment *n* by one from 71 until at least 80% power was obtained, the stopping point would be *n* = 72. However, when the seed is modified, power decreases in this case due to the first estimate of power lying above the true value of power. Based on these figures, the determined necessary sample size may vary from 72 to 79, depending on the seed value and starting sample size chosen. The situation is problematic because increasing sample size should never appear to decrease statistical power. There is a specific sample size that leads the true value of power to increment from below 80% to 80% or above.

A methodologist should not base a sample size determination simply on a usual number of replications (*r* = 1,000) knowing that they may actually be over- or under-estimating their power by 2.6% (i.e. 0.80 ± 2*1.26%), and thus under- or over-estimating

the necessary sample size, respectively. Once the methodologist has found a sample size that appears to provide adequate power with a low number of replications, he or she can determine the number of replications necessary to state the true level of power with a desired level of confidence. For example, suppose we would like to be confident that the power estimate obtained is no more than .0025, or .25% different from the true power estimate, giving a 95% confidence interval for power that is .5% wide, as opposed to 5% wide with r = 1,000. This would assure that rounding to the nearest hundredth in the power estimate would not lead to an erroneous conclusion. Determination of the number of replications necessary begins with the following formula:

$$w_{pi} = 4 * s.e.(power) = 4 * \sqrt{\frac{\pi(1-\pi)}{r}} \tag{2.30}$$

Here, w_{pi} is the desired width of the 95% confidence interval for power (.005) in the current example. The multiplier, 4, is included to reflect the fact that ± 2 standard errors give the appropriate width (w) of a 95% confidence interval, approximately. The subscript *pi* (for *power interval*) is added to avoid confusing this value with w, a symbol others have used to refer to the desired width of a confidence interval for a sample statistic (Kelley, 2007). Now, solving for r gives:

$$r_{necessary} = \left[\frac{4}{w_{pi}} * \sqrt{\pi(1-\pi)} \right]^2 \tag{2.31}$$

For a desired width of .005 at π = .8, the necessary number of replications is 102,400. Thus, 102,400 replications will allow the determination of the sample size n that will provide adequate power, with 95% confidence that the true power value is within the interval: 79.75% to 80.25%.

In order for the determination of power to be stable across multiple seed values, an empirical power value should have a confidence interval no wider than 1% (e.g. .795 to .805); such that when the value is rounded to two decimal places, it will not be likely to change merely as a function of the seed value (should the analysis be performed again with a different seed). A desired width of .01 at $\pi = .8$, gives 25,600 as the number of necessary replications. It is thus recommended that any methodologist performing an empirical power analysis should carry out a minimum of 25,600 replications in order to avoid inadvertently over- or under-powering the proposed study (Kupzyk, 2009; unpublished manuscript).

Current Study

The primary goal of this study is to determine the conditions under which three approaches for conducting power analyses incorrectly estimate the number of subjects or clusters necessary to achieve sufficient statistical power. Planning of longitudinal cluster randomized trials -- or cluster randomized trials with three or more repeated measures (Spybrook, Raudenbush, Congdon, & Martinez, 2009) -- to evaluate education-based interventions is the context of interest. This design has become an increasingly popular and important research paradigm. Even when treatments are randomly assigned to participants there may be unforeseen differences in outcomes prior to the onset of a study. Repeated measures are necessary to take into account differences between individuals and experimental groups at baseline and to estimate developmental trends.

As discussed in this chapter, there are many different methods for determining power, such as subjects to variables ratios, power to detect model fit, power tables and formulae, the “exemplary dataset” method, and simulation-based approaches. There are

many experimental design factors affecting power, including sample size, effect size, Type I error, group balance, attrition, and duration in a longitudinal study, and each power analysis approach requires different decisions to be made regarding these factors. Each research study has a unique set of conditions that can and should be accurately accounted for in a power analysis. The current study focuses on the interaction of study planning methods and experimental design factors in cluster randomized trials with repeated measures – a design that is becoming more common in education research. The study will attach numerical meaning (change in statistical power) and logistical meaning (change in the number of participants) to the penalties incurred by ignoring violations of common assumptions in power analysis.

It is widely accepted that planning a study that will have sufficient power is critical. Some major funding agencies, such as the National Institutes of Health, which encourage the use of randomized clinical trials, now require power analyses (Thoemmes, MacKinnon, & Reiser, 2010). The Institute of Education Sciences (IES), which is the research division of the United States Department of Education, has developed a "goal" structure for determining types of research projects. Efficacy and replication trials in a limited range of settings (Goal 3) and full-scale evaluations (Goal 4) require power analyses (Institute of Education Sciences, 2011), while grants intended for small-scale research, qualitative research, or developing new measurement tools may not require a power analysis. A current Request for Application (RFA) for intervention research through the National Institute of Health states "studies must be appropriately powered" and asks "is the sample size adequate and is the study adequately powered to detect intervention effects?" (National Institutes of Health, 2011). Power analyses are often

conducted in such a way as to provide imprecise estimates of power, leading to an over- or under-estimation in the sample size necessary to carry out an experiment. For example, power formulae that do not take clustering into account overestimate statistical power because when individuals within a cluster tend to be more alike than individuals across clusters, less independent information is being supplied by each individual (i.e. the design effect). A power analysis that makes this incorrect assumption will provide a sample size estimate that is too low to achieve the desired level of power. An inadequate sample size will likely lead to the inability to claim that the effect of interest is significant, thus wasting the resources spent on an experiment and failing to determine if the insignificant effect is due to insufficient sample size or due to the effect not being truly present in the population. Sampling too many participants, in addition to wasting resources, also has severe consequences, including unnecessarily putting too many participants at risk, withholding effective treatments from control group participants, putting unnecessary burden on participants, and potentially detecting an effect size that is too small to be clinically important. Therefore, it is of great import that the most accurate estimates of necessary sample size possible be determined. The results of this study will be valuable to methodologists and researchers interested in using the most accurate and innovative strategies for power analysis and research design.

Primary Research Questions

The four primary research questions are as follows:

1. To what degree does failure to account for attrition bias estimates of sample size and power?

2. To what degree does failure to account for unequal treatment group sizes bias estimates of sample size and power?
3. To what degree does failure to account for unequal cluster sizes bias estimates of sample size and power?
4. Does the amount of variability attributable to clusters (ICC) affect the magnitude of bias imposed by failing to account for these conditions?

It is hypothesized that failure to account for attrition will lead to the most biased estimates of power. Specifically, power should be greatly overestimated by failing to account for attrition because dropout is known to cause severe losses in efficiency (Verbeke & Lesaffre, 1999). Failure to account for unequal treatment group sizes is also expected to bias power and sample size calculations, although to a lesser degree. Konstantopoulos (2010) found noticeably different power values as treatment conditions deviated from equality, although the study investigated only two-level cross-sectional designs. It is hypothesized that failure to account for unequal cluster sizes will not bias power estimates, except for in the skewed distribution condition. No interaction effects are expected, but they will be investigated. The empirical power and "exemplary dataset" methods are expected to produce equivalent results in every condition because both methods accommodate all the design features considered in this study. Any unanticipated differences will be closely examined to determine if one method is less biased than the other.

CHAPTER III. METHODS AND PROCEDURES

The context of interest in this study is when clusters of participants are assigned to one of two experimental conditions, say treatment or control, within a longitudinal study with two or more intended measurement occasions. The assignment of clusters to condition is often referred to as a *cluster randomized design*, or cluster randomized trial (CRT), because intact social clusters -- classrooms in this context -- are the unit of randomization. Such designs are useful in educational research, as individual assignment within classrooms is often difficult and would be prone to treatment contamination. Outcomes are assumed to be measured at the student level.

Although power analysis may be applied to any planned statistical test, this study is concerned only with tests of group differences in the change that occurs as a result of an intervention. Accurate assessments of power for tests of single-group differences from constants, correlations, single-group change, and non-linear change are important. However, tests of differences in linear change between two experimental groups are more common in modern psychological and educational research, in that differences in quadratic trends across treatment groups are not often tested.

Analytic Model

A pre-post analysis of covariance (ANCOVA, i.e. post-test as the outcome, controlling for baseline) can be implemented when there are at least two measurement occasions, intervals are equal across participants, and there is no clustering of participants. In some cases, simple gain scores (post-pre) have been used as outcomes, although many methodologists would argue that change scores are not reliable measures of change due to the correlation between pre-test and post-test scores (Dimitrov &

Rumrill, 2003). Cronbach and Furby (1970) recommended residualized change scores to account for the correlation by expressing the post-test score as the deviation of the score from the regression line formed by regressing post-test on pre-test. The residualized change score thus partials out the correlation between pre and post. The method is equivalent to ANCOVA (Curran & Muthén, 1999), however, and has the same limitations of any pre-post design that does not account for clustering and assumes equal intervals. Such designs do occur frequently in psychology, but not often in educational studies involving multiple classrooms or schools because of the hierarchical nesting structure introduced by the complex sampling of students within classes, and the common use of more than two measurement occasions.

Hierarchical linear models are well-suited for analyzing group differences in change allowing realistic conditions, such as unequal intervals and individually-varying time points, as well as accounting for dependence of individuals due to complex sampling. While models of change may reasonably be analyzed using a repeated-measures analysis of variance (RM-ANOVA) or a multivariate approach to repeated measures (MANOVA) examining the time by group interaction, there are several advantages to using the HLM framework. First, RM-ANOVA and MANOVA assume independence of cases, whereas HLM can account for clustering, and thus the correlation, of individuals within groups. These other approaches also do not take into account variance in individual intercepts or slopes (Ware, 1985). Consequently, RM-ANOVA and MANOVA analyze only mean changes and treat differences between individuals as error variance. Some of this variance may contain valuable information about the developmental change process (Hertzog & Nesselroade, 2003). Second, RM-ANOVA

and MANOVA assume measurements are obtained at equal intervals and that intervals are equal across individuals. In strictly-controlled trials in laboratory settings, this assumption may be tenable. In educational settings where measurements are taken annually or bi-annually, the variation in time intervals will likely be quite small and very little bias would occur by assuming equal intervals. In field settings involving individuals or families, however, differing amounts of time may elapse between measurement occasions and individuals may be measured on different time frames (i.e. unequal intervals *and* individually-varying intervals). Such instances can occur because a child or teacher is sick on the day of testing, a family cannot be contacted on a certain day, or other extenuating circumstances make it impossible for all participants in a study to be measured on exactly equal intervals. In such cases, assuming equal intervals may severely bias estimates of growth trajectories. HLM allows for individually-varying time points providing an accurate account of individual growth trajectories. Third, general linear model approaches in general delete cases in a listwise manner, meaning that any individual that does not have measurements taken at all time points will be entirely excluded from the analysis, whereas HLM utilizes all available time points from an individual. Finally, HLM has generally been found to have higher power to detect intervention effects than RM-ANOVA (Curran & Muthén, 1999). Simply put, RM-ANOVA is no longer considered a state-of-the-art statistical method, and its use in longitudinal designs should be discouraged.

Choice of Power Analysis Methods

In particular, this study will compare and contrast estimates of statistical power and necessary sample size from power analyses employing a) a formulaic approach using

the Optimal Design software (Spybrook, Raudenbush, Congdon, & Martinez, 2009), b) an “exemplary dataset” approach (e.g. O'Brien & Muller, 1993; Stroup, 2002), and c) an empirical approach through Monte Carlo simulation (Muthén & Muthén, 2002; Littell et al, 2006). These three methods are of interest because they are the most widely applied (the formulaic approach) or the most applicable (“exemplary dataset” or empirical approaches) to power analyses for studies in the context of interest. The “subjects to variables ratio” method will not be considered here because it does not account for effect size, which is an important determinant of power. Such a guideline may be useful for collecting pilot data in the absence of any usable effect size assumptions. Although it is easy to implement, the lack of consideration of power and effect size renders the ratio method incomparable to other sample size determination methods.

The RMSEA model fit method of MacCallum, Browne, and Cai (2006), which extends MacCallum, Browne, and Sugawara (1996) by adding the ability to evaluate differences between models, is not considered because it is concerned with global fit of a structural equation model. The current study is focused on power to detect single parameters, specifically the coefficient reflecting the effectiveness of an intervention. MacCallum et al. is not concerned with power to detect any specific parameter of interest, and thus is not of interest in the current study. The Satorra and Saris (1985) method does determine the power of a likelihood ratio test for a single parameter in a structural equation model. The method is similar to the “exemplary dataset” method (Stroup, 2002) in that two models are compared, one holding the parameter of interest at zero, and difference in chi-square values is the non-centrality parameter, which is used to determine power. The “exemplary dataset” method also uses the non-centrality parameter

associated with a regression coefficient, but the current study is focused on the hierarchical linear modeling framework. To apply the Satorra and Saris method, the user must determine two implied correlation matrices among observed variables in a SEM. The current study is conducted in the context of hierarchical linear models for latent growth where hierarchical effects are parameterized as random effects in a mixed model (see Singer, 1998).

Simulation Conditions

Some elements of the experimental design were not varied due to the fact that they are values that are supplied by the user of power analysis methods. They are typically occurring conditions in educational research situations, and variations do not result in violations of assumptions made by power analysis methods.

Fixed Simulation Conditions

Number of occasions. While the number of measurement occasions has an important effect on power, it will be held constant because the methods of interest are assumed to equally and accurately account for the effect of varying the duration of a study. The difference in power estimates across methods may increase or decrease as the number of measurement occasions in modified, but this study will focus on factors that may be accounted for differently across power analysis methods. The experimental design will emulate studies with four measurement occasions. Educational studies may measure students and classrooms during the fall and spring of two consecutive academic years (e.g. Flannery, Vazsonyi, Liau, Guo, Powell, Atha, Vesterdal, & Embry, 2003), four times during one school year (e.g. Raver, Jones, Li-Grining, Metzger, Champion, & Sardin, 2008), or annually across four years (e.g. Melby & Conger, 1996; Hill, Castellino,

Lansford, Nowlin, Dodge, Bates, & Pettit, 2004). A meta-analysis of published articles investigating the effects of parental involvement on child achievement in middle school was conducted by Hill and Tyson (2009). Two longitudinal studies were reported and both tracked students for four waves of data collection. Regardless of the time frame, studies on educational interventions often implement approximately four measurement occasions. Pre-post designs are also very common, perhaps more common, but having more than two occasions increases the reliability of slope estimates (Byrne & Crombie, 2003).

Effect size. The effect size in this study is fixed at $d = .50$, which is a medium effect size (Cohen, 1988). Similar to the number of measurement occasions, effect size is known to have an appreciable effect on power. Although varying the effect size may cause the difference in power and sample size between power analysis methods to change, the exact magnitude of the difference is not of specific interest because the magnitude will always depend on a number of other design features as well. The concern in this study is evaluating the conditions under which there is any noticeable and important difference in power estimates across methods. Due to the longitudinal nature of the design, the effect size is defined as the difference in linear change between the two experimental groups divided by the standard deviation of the slope values (Raudenbush & Liu, 2001). The standard deviation of the slope values is set at 1.0 meaning the difference in linear rates of change between the two groups is set at .50 for a medium standardized effect size.

Manipulated Factors

What is of primary interest in this study are the effects of varying experimental design conditions that affect power, and may not be accommodated by some power analysis methods. Specifically, attrition, treatment group balance, cluster size, and intraclass correlation.

Attrition. Some amount of participant dropout should be expected in any longitudinal study involving human participants (Hedeker, Gibbons, & Waternaux, 1999). The rate of attrition (the percentage lost between measurements) and the number of measurement occasions will determine the total effect on power. A four-occasion study which loses 10% of participants between each time point will retain only 70% of the original sample at the final occasion. Regardless of whether the analytic model deletes cases listwise or uses all available data, as attrition rates rise, power decreases. If a power analysis fails to account for the possibility of participant dropout, the planned sample size will be underestimated to some degree.

A condition might occur in which the attrition rate varies across treatment groups. For example, individuals in a treatment group might be expected to be retained at a higher rate (less dropout) than individuals in a control group. The opposite could also occur if treatment participants improve so much that they feel they no longer need to participate. If these conditions are hypothesized, the power analysis should take the differential attrition rates into account. A differential attrition condition will not be investigated in this study, however, because most power software programs cannot account for any attrition (with the exception being the RMASS2 program; Hedeker et al,

1999), much less differential attrition, and comparisons between software, empirical, and exemplary methods would not be meaningful.

Three attrition conditions will be implemented in this study: 0%, 5%, and 15% per interval. With four measurement occasions, the 5% and 15% per interval conditions will result in total dropout rates of 15% and 45%, respectively, by the fourth occasion. Raver, Jones, Li-Grining, Metzger, Champion, and Sardin (2008) reported 16% attrition from fall to spring among children in a study involving 35 Head-Start preschool classrooms. In a study on the quality of parent-child interactions, Van Doesum, Riksen-Walraven, Hosman, and Hoefnagels (2008) observed a 16% attrition rate over 6 months. The rates proposed are therefore reasonable and conservative estimates of attrition between time points. It is assumed that once a participant drops out of a study, they will not return.

Treatment group balance. Many power software packages that account for clustering in longitudinal studies assume experimental conditions are assigned to clusters or individuals in equal proportions. Konstantopoulos (2010) tested the effects of treatment group imbalance in two-level cross-sectional designs, finding decreasing power as group imbalance increased. Under extreme imbalance, where one group is 15 times the size of the other (94% to 6%), power was found to be reduced by up to 50% in some conditions. The size of the imbalance and the intraclass correlation were found to affect power. Equal assignment optimizes statistical power, but may not be feasible if the cost associated with measuring units in the treatment condition is much higher than in the control condition. Therefore, the effect on power should be assessed to determine the importance of accounting for the imbalance.

Although it may seem unlikely that someone planning a study would knowingly create an imbalance in experimental conditions, such situations do occur and may be inevitable in some cases. In early childhood education, the effectiveness of the Head Start program has been an important area of research in recent years. In a review of seven major studies designed to assess or increase school-readiness, Early et al (2007) reported that, on average, only approximately one-third of classrooms in the studies were Head Start classrooms. Due to the lack of randomization of conditions to classrooms, the Head Start example is a quasi-experimental situation, which is arguably a much more likely setting for unequal group sizes than randomized experiments (Laczo, Sackett, Bobko, & Cortina, 2005). However, the cost of applying an intervention may lead those conducting randomized experiments to create an imbalance as well. If there is little cost involved in measuring control classrooms, a higher proportion may be sampled, relative to treatment classrooms. Conversely, if confidence in an intervention is high, we may wish to extend the benefits of the intervention to as many classrooms as possible and assign the treatment condition to a higher proportion of classrooms, relative to control classrooms.

Another situation in which a researcher might create an imbalance between treatment groups is when the treatment itself dictates the size of the groups. Studies of classroom size, for example, may be imbalanced by design because all classrooms in one condition are smaller or larger than all classrooms in another condition (e.g. Nye, Hedges, & Konstantopoulos, 2000).

Two conditions of treatment group balance will be implemented in this study: 50% of clusters in each condition, and 70% of clusters in the treatment condition (30% control). Which experimental condition receives the higher proportion of units, treatment

or control, is not of substantive interest in this study. An imbalance in favor of either condition leads to a decrease in statistical power. In the previously mentioned meta-analysis conducted by Hill and Tyson (2009), five intervention studies were found where the percent of individuals in the treatment group ranged from 47% to 58%. In the school-readiness studies reviewed by Early et al. (2007), an average of 36% of were Head Start classrooms. The range of Head Start percentages was 9% to 100%. This evidence indicates that 60% and even 70% of units in one experimental, or quasi-experimental, condition is not an unreasonable or uncommon situation.

Cluster size. One may envision studies that combine classrooms from both rural and urban settings, or large and small towns, where classroom sizes will likely be quite different. Much research has been conducted on the effects of class size on outcomes of interest. In order to generalize to a large portion of educational settings, rural, suburban, and urban classrooms should be observed, in order to gain a representative sample of the population of interest. Such sampling will result in a wide range of classroom sizes. Hammarberg and Hagekull (2002) reviewed literature on classroom size and conducted a study investigating the effect of classroom size on teachers' perceived control of child behavior. The study included 22 rural, suburban, and urban pre-school classrooms ranging in size from 7 to 26 students, meaning cluster sizes are often unequal.

Power software programs that account for clustering assume equal cluster sizes. Although this condition often comes up in educational research, the effect of unequal cluster sizes on power is not typically accounted for. Candel, Van Breukelen, Kotova, and Berger (2008) investigated the effect of varying cluster sizes in multilevel studies and found that unequal clusters produce higher effect variances, and thus less efficiency. The

authors found that the loss of efficiency could be compensated for by increasing the number of clusters by 18%. The finding, however, is dependent on the beginning sample size and the size of the effect. This study will examine if, and under what conditions, unequal cluster size impacts statistical power.

Three conditions of cluster size will be implemented in this study. First, an equal size condition will keep cluster sizes constant at 20 individuals per cluster, representing an average-sized classroom in a public school. Second, cluster sizes will vary in a normally-distributed fashion, with an average of 20 individuals per cluster and standard deviation of 2. This results in a range of classroom sizes between approximately 15 and 25. Third, a negatively skewed distribution of classroom sizes will be implemented, as was the case in the Hammarberg and Hagekull (2002) study. This reflects a situation in which all classroom sizes are expected to be of size 20, which is the mode, but there are more classrooms on the lower end of the distribution than on the higher end. In other words, one might expect there to be a sizeable portion of classrooms with 15-20 individuals, but few with higher than 20. While the overall number of classes is held constant, there are more class sizes that are smaller than expected than larger, so the overall number of students is smaller than expected. A positively skewed distribution of classroom sizes will not be assessed, because a situation in which a higher proportion of larger classrooms than expected is obtained is less realistic than ending up with lower class sizes.

Intraclass Correlation. The intraclass correlation, which is accommodated by power software programs for multilevel models, will be set at three levels -- .05, .10, and .15 -- to determine if the magnitude of the difference between power analysis methods

varies depending on the amount of variability in estimated individual trajectories that is attributable to clusters.

Summary of Experimental Design

The four factors described for this simulation study result in a 3x2x3x3 design. Three levels of attrition (0%, 15%, and 45%), two levels of experimental group balance (50/50 and 70/30), three levels of cluster size balance (equal cluster sizes, normally distributed cluster sizes, and a skewed distribution of cluster sizes), and three levels of intraclass correlation (.05, .10, and .15) form the 54 cells of the design. The number of measurement occasions in this study is fixed at four. The design is depicted in Figure 5.

		Treatment Group Balance					
		50/50			70/30		
		Cluster Size			Cluster Size		
ICC	Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
.05	0%	OD: J=16	OD: J=16	OD: J=16	OD: J=16	OD: J=16	OD: J=16
	5%	OD: J=16 Adj: J=18	OD: J=16 Adj: J=18	OD: J=16 Adj: J=18	OD: J=16 Adj: J=18	OD: J=16 Adj: J=18	OD: J=16 Adj: J=18
	15%	OD: J=16 Adj: J=23	OD: J=16 Adj: J=23	OD: J=16 Adj: J=23	OD: J=16 Adj: J=23	OD: J=16 Adj: J=23	OD: J=16 Adj: J=23
.10	0%	OD: J=22	OD: J=22	OD: J=22	OD: J=22	OD: J=22	OD: J=22
	5%	OD: J=22 Adj: J=25	OD: J=22 Adj: J=25	OD: J=22 Adj: J=25	OD: J=22 Adj: J=25	OD: J=22 Adj: J=25	OD: J=22 Adj: J=25
	15%	OD: J=22 Adj: J=32	OD: J=22 Adj: J=32	OD: J=22 Adj: J=32	OD: J=22 Adj: J=32	OD: J=22 Adj: J=32	OD: J=22 Adj: J=32
.15	0%	OD: J=28	OD: J=28	OD: J=28	OD: J=28	OD: J=28	OD: J=28
	5%	OD: J=28 Adj: J=32	OD: J=28 Adj: J=32	OD: J=28 Adj: J=32	OD: J=28 Adj: J=32	OD: J=28 Adj: J=32	OD: J=28 Adj: J=32
	15%	OD: J=28 Adj: J=41	OD: J=28 Adj: J=41	OD: J=28 Adj: J=41	OD: J=28 Adj: J=41	OD: J=28 Adj: J=41	OD: J=28 Adj: J=41

Figure 5. Experimental design. **Note.** *OD* refers to a projected sample size reported by the Optimal Design software where *J* is the number of clusters. *Adj* indicates the adjusted number of clusters needed by accounting for the rate of attrition.

Procedures

The study will begin by using Optimal Design to determine the number of clusters it deems necessary to achieve 80% power in each cell of the design. In cells where the treatment groups are imbalanced, balanced groups will be assumed. Where cluster sizes vary, the average within-cluster sample size will be assumed. Three numbers of clusters labeled "OD: J=" are shown in Figure 5, varying only on ICC. The number of clusters shown will be used for all cells with each respective level of ICC. As a follow-up analysis under conditions of attrition, the percent of the sample that is projected to be lost by the end of the study will be added to the baseline sample size, labeled "Adj: J=" in Figure 5. The number of clusters obtained using Optimal Design will then be used in an empirical power analysis and an exemplary dataset power analysis to determine the empirical power value obtained under the conditions of each cell. The resulting power and sample size point estimates will show the amount of bias caused by making simplifying assumptions with power analysis software.

In each cell of the design, an empirical power analysis and an exemplary dataset power analysis will then be carried out to determine the number of clusters necessary to achieve sufficient power under the conditions of each cell. Comparing these values to the necessary sample size values reported by Optimal Design will show the amount of bias in the number of necessary clusters caused by making simplifying assumptions with power analysis software. Although the bias in power will be previously determined, assessing

bias in sample size will make the observed effects of simplifying assumptions more interpretable and meaningful.

Software for Empirical Power and Exemplary Dataset Power

There are multiple software packages and computing environments that are available for performing empirical power analyses. In fact, any programming environment that supports data generation and data management can be used as well as packages that have a direct simulation procedure imbedded (e.g. Mplus, LISREL, and EQS). However, reviewing all the possible software programs for conducting an empirical power analyses is not a goal of the current study. All data generation, data manipulation, analysis, and tabulation of results for the empirical and exemplary dataset power analysis methods will be carried out using SAS. The SAS syntax for the empirical power analysis used in one cell of the design is shown in Appendix A. Following Kupzyk (2009), r = 26,000 replications per each cell in Figure 5 will be carried out in each empirical power analysis. This will ensure stable estimates of power, and thus necessary sample size. Appendix B provides the exemplary dataset syntax for all cells within one level of ICC.

The SAS data step and Interactive Matrix Language (IML) are extremely flexible tools for generating data (using the RAND function for generating random numbers), manipulating datasets, and collating analysis results. This SAS functionality allows the user to specify proportions and patterns of missing values, treatment group balance, and cluster sizes. In addition, many procedures in SAS, including MIXED, which will be used for analyzing each empirical and exemplary dataset, have an Output Delivery System (ODS OUTPUT), through which the tests of fixed effects in each analysis will be

saved. In an empirical power analysis, after all the simulated datasets have been analyzed using PROC MIXED, the p-values associated with the test of the time by group interaction will be transformed into a column of decisions about significance. If the p-value is less than .05 the result is 1 for a significant interaction. If the p-value is greater than .05 the result is 0 for a non-significant interaction. The average of the column of decisions is the empirical power rate, as it represents the proportion of replications resulting in finding the effect of interest to be significant.

Degrees of Freedom in Empirical and Exemplary Methods

There are five methods available in PROC MIXED for determining the denominator degrees of freedom (DDF) for testing the significance of fixed effects. Although results have been mixed regarding their accuracy (Schaalje, McBride, & Fellingham, 2001), Satterthwaite and Kenward-Roger DDF methods are generally considered superior to residual, containment, and between-within methods. While all five methods are available in an empirical power analysis because a complete dataset (in terms of both fixed effects plus sampling error and individual differences variance) is generated in every replication, Satterthwaite and Kenward-Roger DDF cannot be directly implemented in the exemplary dataset method because they are approximate methods, which means they are based on approximations of the variance components. Variance components are fixed values in the exemplary dataset method, so no estimation of the variance-covariance matrix of fixed effects takes place. Therefore, the *between-within* DDF method will be used in this study because it partitions the degrees of freedom into within- and between-participants portions is the most accepted of the three exact DDF methods (Schaalje, McBride, & Fellingham, 2001). The between-within method will be

implemented in both the empirical and exemplary dataset method to ensure their comparability.

Optimal Design

There are several formulaic power programs available that compute power for longitudinal, multilevel analyses, such as Power in Two Level Designs (PINT; Snijders & Bosker, 1993) and RMASS2 (Hedeker, Gibbons, & Waternaux, 1999). In particular, Optimal Design 2.0 (Spybrook, Raudenbush, Congdon, & Martinez, 2009) provides power and sample size estimation in cross-sectional and longitudinal designs that assign treatments at either the individual- or cluster-level, with person- or group-level outcomes. Optimal Design is a flexible and user-friendly program, so the formulaic power method will rely on power and sample size estimates supplied by Optimal Design. The program allows the user to specify cluster size, alpha, the number of measurement occasions, and the intra-class correlation coefficient, which is essential to assessing power in multilevel designs. Optimal Design assumes all designs are balanced and have complete data. No recommendations are made for accounting for attrition.

CHAPTER IV. RESULTS

In order to verify the simulation accuracy in the empirical power analyses, model solutions were checked for indications of convergence problems. Out of the 54 cells of the design, four cells completed 25,599 of the 25,600 intended replications, for a convergence rate of 99.9997%. Each of these four cells that had one replication in which the model did not converge was in the .15 ICC condition. Across the design there were no replications that produced zero values for all five variance components in the model. The person-level slope variance or slope variance were never estimated to be zero. The person-level intercept variance and cluster-level intercept and slope variances, however, did return some zero variance estimates, which are inadmissible because they indicate that the variance was estimated to be negative. Only three replications had zero estimates for the person-level intercept variance, while within-cell percentages of replications with zero estimates for the cluster-level intercept ranged from 0% to .13% (average = .03%). Within-cell percentages of replications with zero cluster-level slope variances ranged from 0% to 11% (average = 3%). The convergence rates for specific variance components were lower at lower ICC conditions. When the ICC is lower, cluster-level slope variances are closer to zero, increasing the likelihood of a zero variance estimate. In these cases, cluster-level variance components could be fixed at zero because there is little or no variance attributable to clusters. The convergence rate for the fixed effects and the high rates of variance component convergence across most cells of the design indicate that non-convergence was not problematic in the empirical simulation portion of this study. Tables C.1-C.6 in Appendix C compare average variance components and parameter estimates to the population values used to simulate the data across all cells of

the design. The tables show the simulation to be very accurate. The average time*group interaction coefficient, for which the population value was .5, never deviated more than .003 across the 54 cells. Similar or small estimates were found for the ICC and all variance components.

Optimal Design was first used to determine the number of clusters necessary to achieve 80% power in each cell of the design. There are three cells of the design (varying only on ICC) where all three methods should agree exactly in their estimates of power. When there is no attrition, treatment group sizes and cluster sizes are equal and balanced. In these cells, the assumptions made by Optimal Design are met. As the current study was being carried out it became apparent that the empirical and exemplary dataset methods were quite similar but there was a sizeable difference in power (3%-5%) between both empirical methods and the formulaic approach implemented in Optimal Design. Upon programming the formulae provided in the Optimal Design manual into SAS, the difference was found to be due to the degrees of freedom assumed for the test of the time by group interaction. As assignment to condition occurs at the cluster level, Optimal Design (and the HLM software program, which is produced by several of the same authors; Raudenbush, Bryk, & Congdon (2004)) assumes J-2 degrees of freedom, where J is the number of clusters. When the hierarchical linear model used in this study is implemented in SAS, however, the degrees of freedom value is much higher. Using the between-within method, the degrees of freedom is:

$$df_{BW} = (J * n * t) - J - 2 \tag{4.1}$$

where *n* is the cluster size and *t* is the number of measurement occasions. This is based on the intervention effect being tested as a cross-level interaction. For example, when there

are 16 Clusters of size 20 across 4 time points, Optimal Design uses *df* = 16 - 2 = 14 for the test of significance, whereas the MIXED procedure in SAS uses *df* = 16*20*4-16-2 = 1262. The result is an increase in estimated power using both the empirical and exemplary dataset methods over the formulaic approach. Consequently, formulaic power was calculated and reported based on Equation 4.1.

Initial Results

Figure 6 shows a plot of the power estimates obtained with the empirical and exemplary dataset methods in the 18 cells of the .05 ICC condition. Optimal Design reported 16 clusters were necessary to achieve over 80% power, which exceeds the 80% criterion by 2.7%. Although power based on J = 16 clusters exceeds 80%, power with J = 15 clusters is below 80% (79%), so J = 16 is used as the referent value. In the cell where all assumptions were met, namely 0% attrition, equal cluster sizes, and equal treatment group balance (top left corner of Figure 6, labeled with a star), both the empirical and exemplary dataset methods estimate power to be 4%-5% higher than the formulaic approach as currently implemented in Optimal Design. With 16 clusters, the horizontal line labeled “df = J-2” is where the empirical and exemplary power estimates in this cell were expected to fall. The upper horizontal line is where the formulaic method would have projected power to be if the between-within degrees of freedom method was assumed.

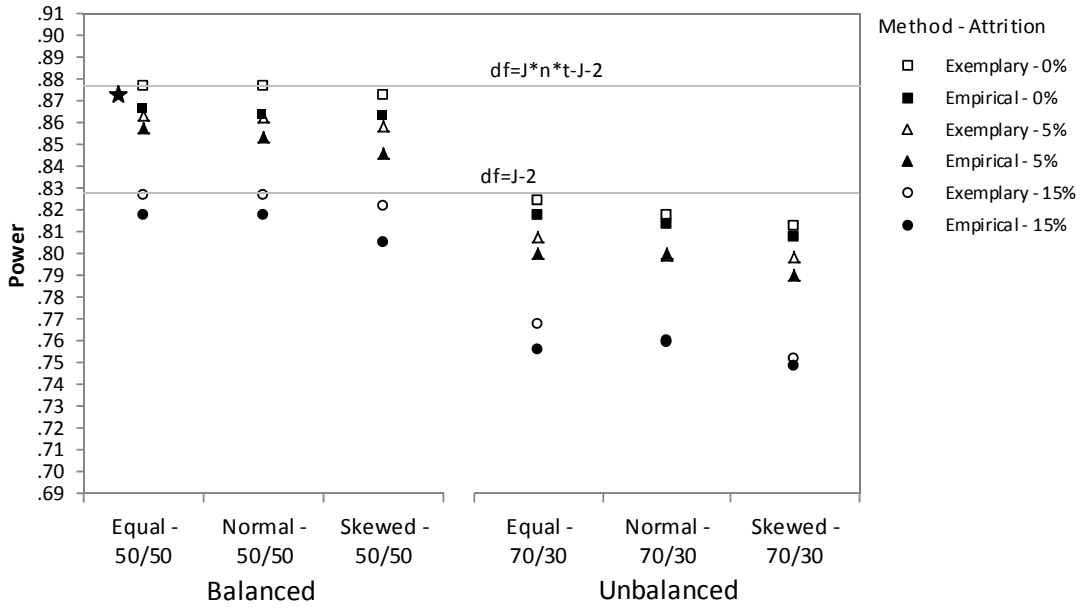

Figure 6. Power estimates under the .05 ICC condition (J = 16) paneled by intervention group balance (50/50 versus 70/30).

The exemplary dataset method produces a power estimate that is nearly identical to the between-within formulaic method line and the empirical power estimate is within 1%, which would likely decrease with an increased number of replications. Figures 7 and 8 present the results for the .10 and .15 ICC conditions, respectively.

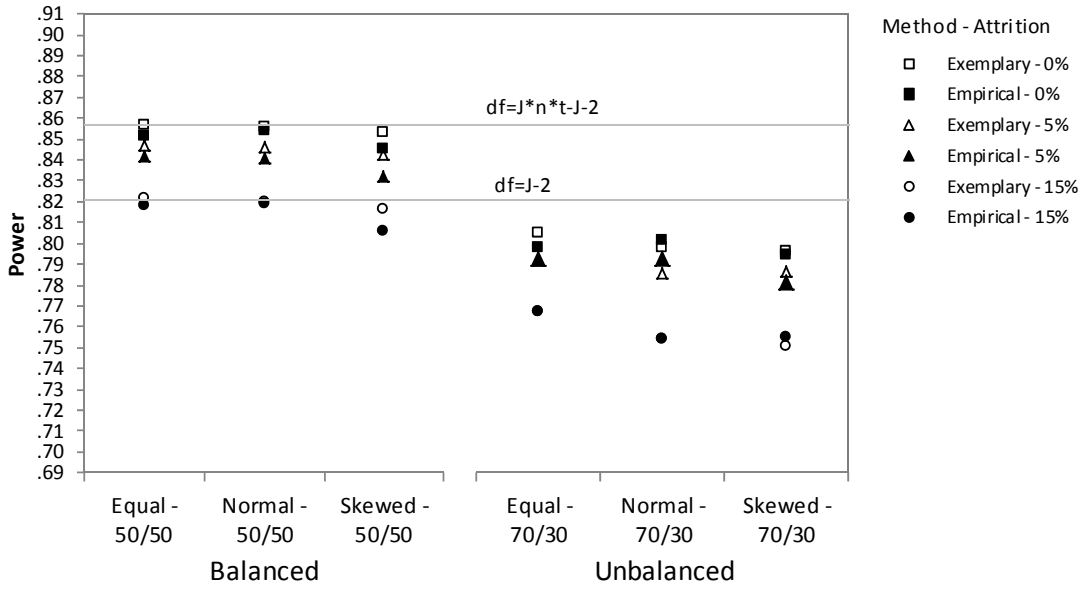

Figure 7. Power estimates under the .10 ICC condition (J = 22) paneled by intervention group balance (50/50 versus 70/30).

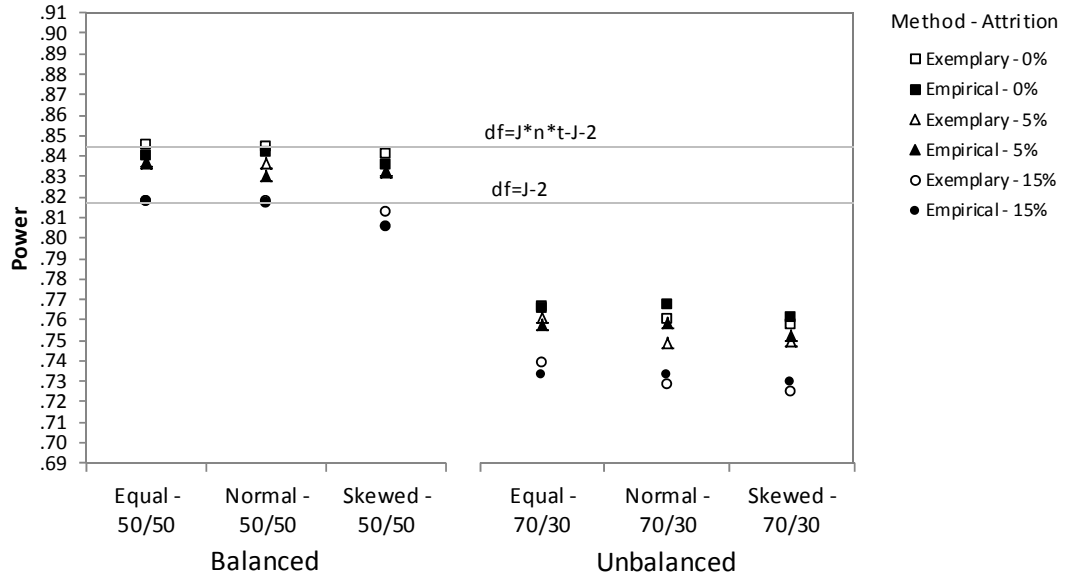

Figure 8. Power estimates under the .15 ICC condition (J = 28) paneled by intervention group balance (50/50 versus 70/30).

Upon inspecting Figures 6-8, several trends are apparent. First, cluster balance does not appear to have an appreciable effect on power. Power does decrease as attrition increases (0%, 5%, and 15% per interval). Across all other cells of the design, power decreases by 1.1% as a result of a 5% per-interval attrition rate and by an average of 4.1% as a result of a 15% attrition rate. Power decreases as treatment condition balance deviates from equality (.5 to .7) by an average of 6.7%. The effect on power as a result of attrition varies little across the experimental factors. The effect of treatment group imbalance, however, does interact with the levels of cluster size, attrition rate, and ICC. Specifically, higher declines in power as a result of treatment group imbalance are observed at higher levels of attrition and ICC, and when cluster sizes are unequal. The effect does not vary depending on whether cluster sizes are normally distributed or skewed.

Figures 6 to 8 show that the empirical power analysis method produced power estimates slightly lower than the exemplary dataset method in many, but not all, conditions. In the first cell of the design the empirical estimate was 1.1% below the exemplary estimate (the star shown in the top left corner of Figure 6). Differences of approximately the same magnitude were observed in several cells of the design. Empirical analyses were performed with 25,600 replications in each cell of the design. In order to determine if the difference is due to sampling variation in the empirical power analyses, a second empirical analysis was performed for the first cell using 50,000 replications. The resulting power estimate was 0.9% below the exemplary estimate, compared to 1.1% below using the smaller number of replications. The empirical estimates appear sufficiently close to exemplary estimates to assume the differences

found are due to sampling variations and as the number of replications increases the two methods will converge.

The power estimates obtained using the empirical and exemplary dataset methods in the conditions where all assumptions are met are well above 80%. In order to address the primary research questions and better interpret the biases injected by failing to account for violations of assumptions, the design was carried out again using the number of clusters necessary to achieve sufficient power according to the formulaic approach using between-within degrees of freedom. To achieve 80% power, 13, 19, and 25 clusters are needed in the .05, .10, and .15 ICC conditions, respectively. Only the exemplary dataset method was performed in this part of the analysis. The resulting power estimates from this analyses are shown in Figures 9 to 11. The exact observed power estimates from the exemplary dataset method are presented in Table 1.

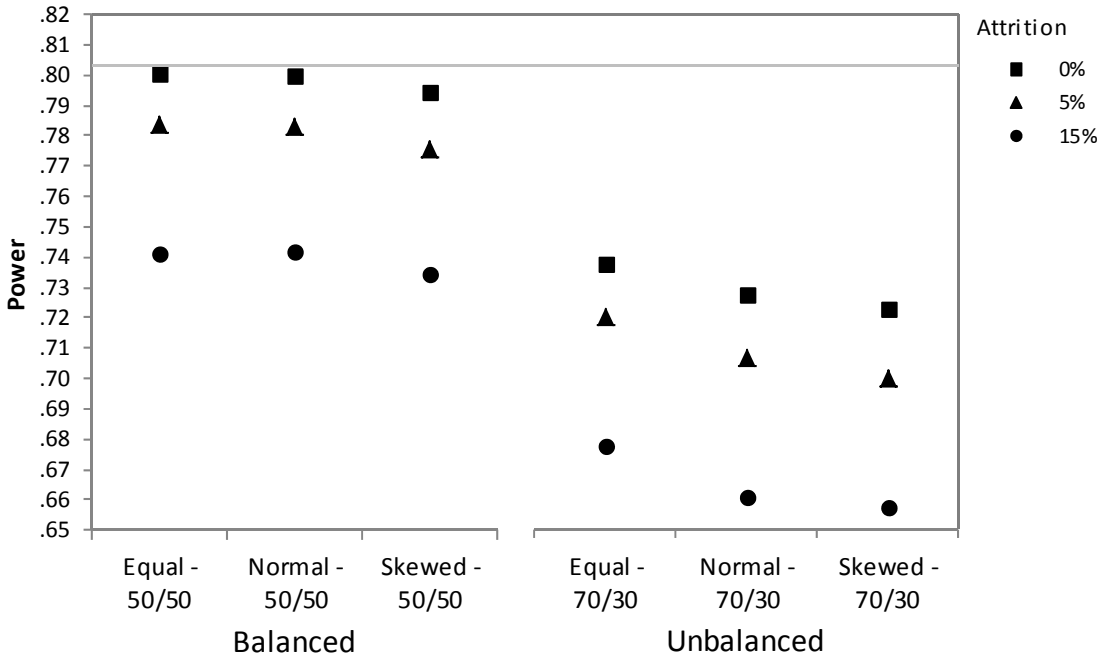

Figure 9. Exemplary power estimates under the .05 ICC condition (J = 13) paneled by intervention group balance (50/50 versus 70/30).

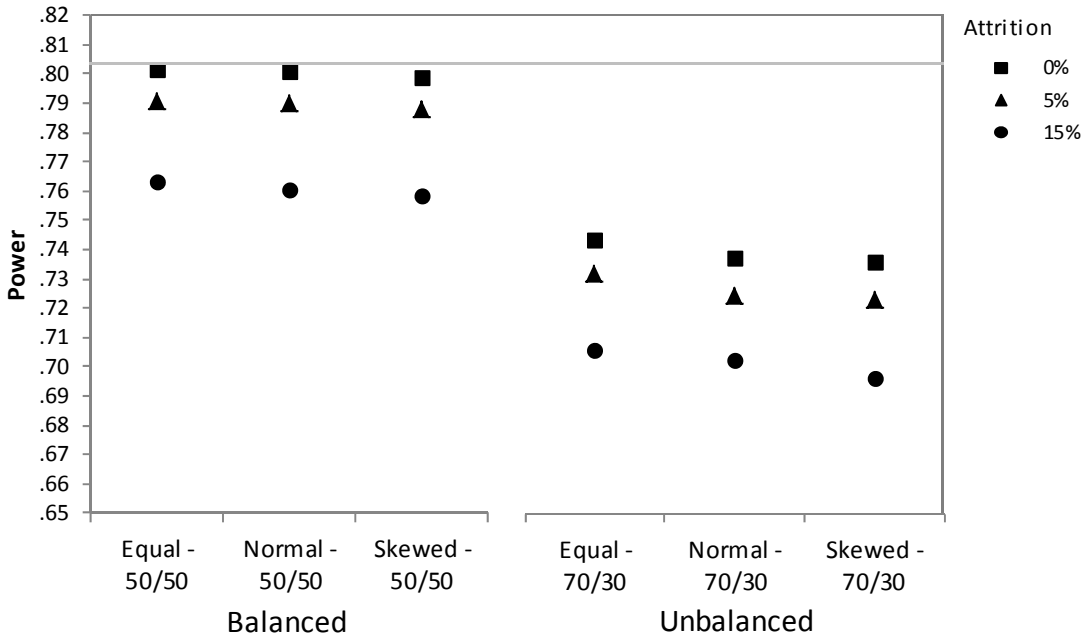

Figure 10. Exemplary power estimates under the .10 ICC condition (J = 19) paneled by intervention group balance (50/50 versus 70/30).

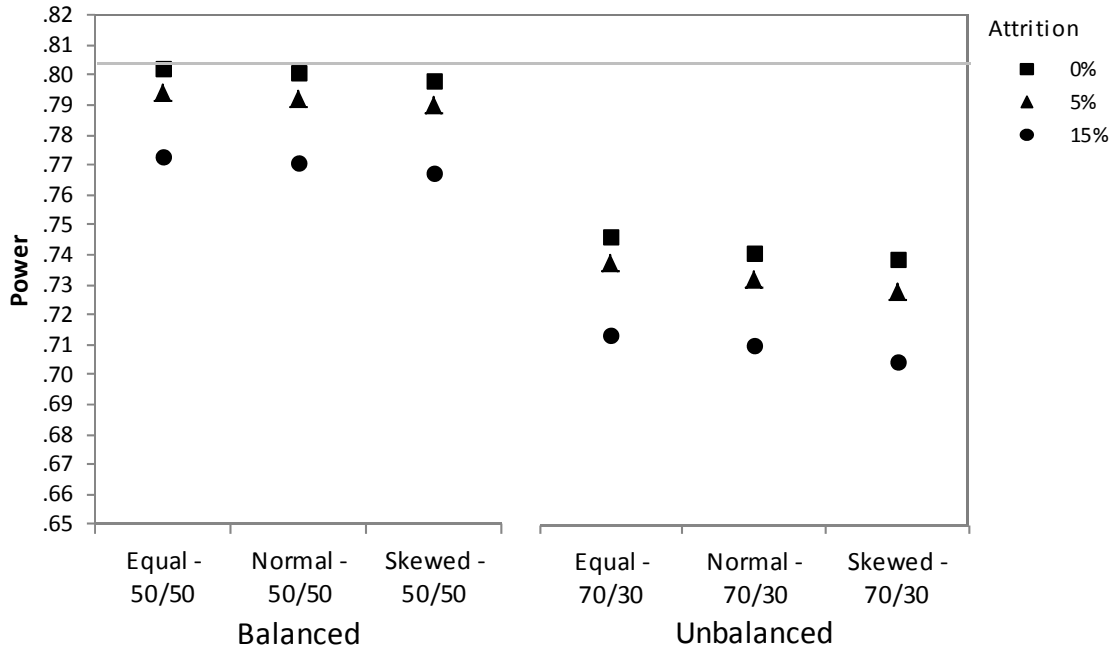

Figure 11. Exemplary power estimates under the .15 ICC condition (J = 25) paneled by intervention group balance (50/50 versus 70/30).

Table 1. *Exemplary dataset power estimates for each cell of the design.*

		Treatment Balance (Cluster Balance)					
		Balanced			Unbalanced		
ICC	% Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	.80081	.80014	.79449	.73777	.72741	.72282
.05	5%	.78387	.78305	.77574	.72019	.70682	.70001
	15%	.74109	.74162	.73421	.67773	.66105	.65759
	0%	.80178	.80097	.79878	.74348	.73735	.73629
.10	5%	.79061	.79021	.78771	.73211	.72449	.72328
	15%	.76293	.76033	.75824	.70607	.70266	.69609
	0%	.80210	.80107	.79816	.74630	.74106	.73885
.15	5%	.79380	.79233	.78980	.73745	.73178	.72746
	15%	.77312	.77074	.76734	.71359	.71009	.70471

Research Question 1: To What Degree Does Failure to Account for Attrition Bias Estimates of Sample Size and Power?

Average power estimates across the three ICC conditions were calculated to assess the effect of failing to account for attrition. The average power estimates obtained using the exemplary dataset method and the decreases observed as a result of attrition are presented in Table 2. The result of 5% attrition per interval is a decrease in power ranging from 1.2% to 1.6%, with an average of 1.3% across the 6 treatment balance by cluster balance conditions. Higher losses were seen in the 70% treatment condition under normal and skewed cluster distributions. The decrease in power as a result of 15% attrition per interval ranges from 4.3% to 4.7%, with an average of 4.4%.

Table 2. *Decreases in power as a result of attrition.*

	Treatment Balance (Cluster Balance)					
	Balanced			Unbalanced		
% Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
0%	0.80156	0.80073	0.79714	0.74252	0.73527	0.73265
5%	0.78942	0.78853	0.78442	0.72992	0.72103	0.71692
15%	0.75905	0.75756	0.75326	0.69913	0.69127	0.68613
0%-5%	0.01214	0.01220	0.01272	0.01260	0.01424	0.01573
0%-15%	0.04252	0.04316	0.04388	0.04339	0.04401	0.04652

In the .05, .10, and .15 ICC conditions, respectively, 13, 19, and 25 clusters were found to be needed to achieve sufficient power. As attrition is known to cause a reduction in power, some account for this effect could be made without exact estimates of the reduction. Researchers might oversample by the percent of cases that are expected to be

lost by the conclusion of a study (Rigby & Vail, 1998). With 5% attrition per interval, 15% of cases will be lost by the end of a study and with 15% attrition per interval, 45% will be lost. The number of clusters needed using this method is (Jn+Jn*proportion lost)/n, or the total sample size plus the proportion lost, in cluster size units. Oversampling by 15% and 45%, respectively, in the .05 ICC condition results in 15 and 19 initial clusters needed, as opposed to 13 clusters assuming no attrition. In the .10 ICC condition, 22 and 28 clusters are needed, and in the .15 ICC condition, 29 and 36 clusters are needed. The power estimates obtained using the exemplary dataset method under the 50% treatment group, equal cluster sizes conditions, where power under no attrition is approximately equal to 80% are shown in Table 3.

Table 3. *Power estimates from adding clusters to account for attrition.*

	ICC		
Overall % Attrition	0.05	0.10	0.15
15%	.83961	.84704	.84977
45%	.88187	.90161	.90204

Oversampling by 15% results in power of 84% to 85%. When 45% attrition is accounted for by oversampling, power is estimated to be 89% to 90%. The effects of too much statistical power are wasting resources and putting more participants at risk than absolutely necessary. Another possible heuristic would be to oversample by the percentage of total information lost, or the percentage of level-1 units lost. The 15% and 45% overall attrition levels cause a loss of 7.5% and 22.5% of total level-1 units, respectively. In other words, up to 45% of individuals have missing level-1 units, which are the time-varying dependent variable scores, by the final measurement occasion.

Oversampling by these percentages led to power estimates closer to, but still larger than, 80%.

Research Questions 2. To What Degree Does Failure to Account for Unequal Treatment Group Sizes Bias Estimates of Sample Size and Power?

The power estimate in each cell of the design under the 50/50 treatment split condition, where half of clusters were assigned to the treatment status and half to control, was compared to estimated power in the equivalent cell of the 70/30 condition. That is, power values in each cell of the first three columns of Table 1 are compared to the equivalent cells of the last three columns. Across all 27 conditions, varying on ICC, attrition rate, and cluster size balance, the declines in power due to treatment group imbalance ranged from 5.6% to 8.1%, with an average decline of 6.4%. The average decline due to treatment imbalance was 5.9% in the cluster balance condition where cluster sizes were equal and 6.6% in both the normally-distributed cluster sizes and skewed cluster sizes conditions.

Research Questions 3. To What Degree Does Failure to Account for Unequal Cluster Sizes Bias Estimates of Sample Size and Power?

At each respective level of ICC, attrition, and treatment balance combination in Table 1, the power estimate in the equal cluster sizes condition was compared to the power estimates from the normally distributed and negatively skewed cluster size conditions, found to the first and second cells to the right in Table 1. This resulted in two difference values for each of the 18 ICC by attrition by treatment balance conditions. Declines in power ranged from 0% to 2%, with an average decline of 0.45% in the normally distributed condition and an average decline of 0.85% in the negatively skewed

condition. The average decline due to either normal or skewed distributions was 0.3% in the 50/50 treatment balance condition and 1.0% in the 70/30 condition. Overall, the average decline due to unequal cluster sizes was 0.65%.

Research Questions 4. Does the Amount of Variability Attributable to Clusters (ICC) Affect the Magnitude of Bias Imposed by Failing to Account for These Conditions?

Under violations of each of the three assumptions investigated, higher losses in power were seen at the lowest level of ICC: .05. As can be seen in Figures 9 to 11, the highest decline as a result of attrition, unequal treatment groups, and unequal cluster sizes is observed where the amount of variance attributable to clusters is the lowest. The negative bias in power due to violations of assumptions at each level of ICC is presented in Table 4.

Table 4. *Average declines in power due to violations at each level of ICC.*

	Condition				
ICC	5 % Attrition	10 % Attrition	70/30	Normal Sizes	Skewed Sizes
.05	1.9%	6.2%	7.2%	0.7%	1.3%
.10	1.2%	3.9%	6.1%	0.3%	0.6%
.15	0.9%	3.1%	6.0%	0.3%	0.7%

For each violation in the design, the bias in power is greatest at the lowest level of ICC and decreases as the ICC increases. The increase in power lost as ICC decreases is not linear. The difference in bias between the .15 and .10 ICC levels is much smaller than the difference between the .10 and .05 levels. The bias imposed by failing to account for violations of assumptions is mitigated to some extent by increasing levels of ICC.

Increased Clusters Needed for Sufficient Power

Table 1 showed that ignoring violations of assumptions can decrease power by 10% to 14% in some cases, especially when multiple violations are made. In order to better interpret the meaning of the observed declines in power, additional exemplary dataset analyses were performed in each cell where power did not meet or exceed 80%. Table 5 shows the determined number of clusters that is actually needed to achieve sufficient power under each set of conditions.

Table 5. *Actual number of clusters needed for sufficient power.*

			Treatment Balance (Cluster Balance)					
	%		Balanced			Unbalanced		
ICC	Attrition	(OD)	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	(13)	13	13	14	15	15	15
.05	5%	(13)	14	14	14	16	17	17
	15%	(13)	16	15	16	19	19	19
	0%	(19)	19	19	20	22	23	23
.10	5%	(19)	20	20	20	23	24	25
	15%	(19)	21	21	22	25	25	25
	0%	(25)	25	25	26	29	30	31
.15	5%	(25)	26	26	26	31	32	32
	15%	(25)	27	27	28	32	32	33

The column labeled "(OD)" gives the number of clusters needed, as determined using the formulaic approach of Optimal Design with between-within degrees of freedom, when ignoring violations of assumptions. The increases in the number of clusters range from zero to eight. As an example, Table 5 indicates that if a power analysis is performed using the formulaic approach assuming equal treatment groups and equal cluster sizes

with ICC = .05, 3 more clusters than the reported 13 will be needed to account for 15% attrition per interval. The exact numbers reported here are not generalizable to studies with different numbers of measurement occasions, cluster sizes, or effect sizes. The important point is that failing to account for the design elements considered here has a meaningful effect on the planned sample size. As one would expect, the greater increases tend to occur in conditions where greater losses in power were observed. Upon comparing the increased numbers of clusters to losses in power, however, the largest increase in number of clusters needed did not coincide with the largest loss in power. The largest increases in numbers of clusters needed were generally found at higher levels of ICC and under conditions of treatment group imbalance.

CHAPTER V. DISCUSSION

Longitudinal cluster-randomized trials often encounter conditions that deviate from the simplifying assumptions made by formulaic methods for determining statistical power. The current study was carried out to determine if and to what extent bias in power estimates is imposed by failing to account for such conditions. Specifically, conditions of attrition, unequal treatment group sizes, and unequal cluster sizes were investigated to determine their effect on power and if effects varied as a function of the amount of variability attributable to clusters (ICC). It was hypothesized that each of the violations would lead to decreases in statistical power. No hypotheses were made regarding the magnitude of the bias imposed or if the bias was dependent on the level of ICC.

Discussion of Research Questions

While it may be widely acknowledged that attrition affects power by decreasing the amount of information, precision, and efficiency with which parameters can be estimated (Verbeke & Lesaffre , 1999), the extent of the effect is not precisely known. Attrition was found to negatively bias power estimates by 1.3% where 5% of cases were lost per interval (15% overall). Power decreased by 4.4% when 15% were lost per interval. A common heuristic for dealing with attrition while using formulaic power method is oversampling by the expected amount of total attrition (Overall, Shobaki, Shivakumar, & Steel, 1998). This led to positive bias in power estimates (Table 3), meaning power was too high. In some cases, a 10% positive bias in power estimates was observed. Another heuristic, oversampling by the percentage of total information lost, or the percentage of level-1 units lost, was also explored. The 15% and 45% overall attrition levels cause a loss of 7.5% and 22.5% of total level-1 units, respectively. Oversampling

by these percentages reduced the degree to which the study would be overpowered, but power was still larger than 80%. It is important to note that the bias found using either heuristic varied by the percent of attrition and the level of ICC. Under the planned study conditions, the fact that the exact amount of bias imposed by attrition varied according to the level of treatment group balance and cluster size balance (Table 2) as well as the level of ICC (Table 4) further indicates that an accurate heuristic for the increased sample size needed to account for attrition would depend on more than just the amount of attrition. The entire set of study conditions needs to be considered to properly account for the effect of attrition. Given that the effect of attrition on power varies as a function of several other design elements, it is unreasonable to assume that a single heuristic could accurately account for the change in power as a result of attrition.

It is important to note that the results presented in this study generalize only to analyses that will utilize all available data, as opposed to analyses that delete cases in a listwise fashion. If a researcher intends to analyze data only from the portion of a sample that completes a study (completers only), then the power analysis can be carried out without accounting for attrition as long as the research oversamples by the expected rate of attrition. Say, for example, a power analysis assuming no attrition reports that 200 participants are needed for sufficient power and that the researcher expects an overall attrition rate of 15% by the end of the study. Adding 15% of 200 to the necessary sample size gives 230 participants that should be obtained at the onset of the study. Such a procedure may seem insensible, given that the data that are available on the participants that dropped out of the study before completion can contribute to statistical power, but if inference is to be made only to individuals that complete a full-length intervention, then

including data points from individuals that have dropped out of a study may actually bias the interpretation of an effect and hinder the generalizability of the findings. The current study, however, is focused on studies that intend to generalize to the population of individuals that take part in a study regardless of how many time points they complete. Thus, to properly account for attrition, the power analysis should take into account the total number of data points available from all individuals as well as the entire set of design conditions.

Failure to account for unequal treatment group sizes negatively biased power estimates by an average of 6.4% across all study conditions. The exact amount of bias imposed by treatment imbalance, like the effect of attrition, depends on the level of ICC. The level of attrition or cluster size balance, however, did not substantially moderate the effect of failing to account for treatment balance. Across all conditions, there was a sizeable decline in power, ranging from 5.6% to 8.1%, due to failing to account for imbalance in treatment conditions. The effect of imbalance on power is likely due to the increased standard error of the treatment effect (Konstantopoulos, 2010).

Unequal cluster sizes did not have an appreciable effect on estimates of power. When treatment groups were balanced and cluster sizes were either normally distributed or negatively skewed, power was decreased by less than 1% across all levels of ICC and attrition. Under conditions of treatment imbalance, bias was at or below 1% at higher levels of ICC. Only in the .05 ICC condition where treatment groups were imbalanced, was bias in power above 1%, being closer to 2% for negatively skewed cluster sizes under high rates of attrition. Unequal cluster sizes, therefore, had the smallest effects of any of the three types of violations examined in this study. and is not likely to be of any

consequence. The effect of this violation was probably the smallest because the overall sample size is unaffected and that the distribution of cluster sizes is evenly distributed across treatment groups. Furthermore, the information lost due to smaller clusters is offset by the information gained from clusters that are larger than average. In most cases, researchers that expect cluster sizes to be unequal could safely ignore this condition and assume equal sizes when performing a power analysis.

The amount of variability attributable to clusters had an unexpected effect on the amount of bias imposed by making simplifying assumptions in a power analysis. Although the differences are not always substantial, in all cases higher levels of bias were observed at lower levels of ICC. The effect was not linear, in that the change in bias associated with changing ICC from .05 to .10 was much greater than the change in bias from changing ICC from .10 to .15. The result that lower biases were observed as ICC increased was surprising because increasing ICC is known to decrease power. Yet it was found that when more variability is attributable to the individual (low ICC) the biases imposed by attrition, unequal treatment groups, and unequal cluster sizes are greater. In retrospect, the finding is intuitive because outcome scores and change are individual-level entities, and as more variability is attributable to clusters, then more is known about individuals by their cluster membership. As the ICC decreases, more weight is placed on the individual than on the cluster, because more independent information is available from each individual. As the ICC increased, however, it was also found that more clusters were needed to make up for biases imposed in order to achieve sufficient sample size. Although biases in power were less at higher levels of ICC, more clusters, and thus more overall participants, were needed to make up for those biases

The biases imposed by unequal cluster sizes are not a cause for concern in performing power analyses. Attrition and treatment group imbalance, however, do impose sizeable biases in power estimates across all design conditions investigated. Decreases in power in this study that were approximately 5% (between 4% and 6%) were associated with needing an extra two to four clusters to achieve sufficient power. Although there is no test statistic available to say that those biases are significant, a 5% drop in power is enough to know that if these conditions of attrition or treatment imbalance are anticipated, an analyst needs to take those into account when planning and performing a power analysis. If statistical power drops more than .5% below 80% then sample size and power are insufficient. The established standard for sufficient power in the social and behavioral sciences is 80%, not 79%, and that will not soon change. Although the probabilities for finding true effects and making Type II errors (1-power; the probability of finding a null effect to be significant) are nearly equal between 80% and 79%, the standards have to be set somewhere, and 80% power with a 20% probability of a Type II error have gained acceptance as the probabilities that researchers and reviewers are comfortable with.

Degrees of Freedom

The inferences made in this study regarding the primary research questions were derived from differences between exact power estimates and those found making simplifying assumptions through a formulaic approach to power estimation. The formulaic approach used, however, is different in one respect from how the formulaic approach is implemented in Optimal Design. It was found that Optimal Design (and the HLM software package) assumes a much lower number for the degrees of freedom, J-2,

when testing the significance of the time by group interaction than the mixed model uses in SAS. Using J-2 as the degrees of freedom emanates from the view that the cluster-level is the level of randomization so the number of clusters is the most important sample size and that is the only sample size that should affect the strictness of tests of significance. The degrees of freedom method implemented in SAS takes into account the total number of level-1 observations using the between-within method for degrees of freedom (J*n*t-J-2). The fact that the overall sample size can change without changing the number of clusters, by varying the size of the clusters or the number of time points, has an important effect on power and is reflected by the increased degrees of freedom implemented in SAS. The result of the increased degrees of freedom was a 4%-5% increase in power when using the exemplary dataset or empirical power analysis methods even when none of the simplifying assumptions made by the formulaic approach were violated. The higher degrees of freedom implemented in the empirical and exemplary approaches led to higher power because critical values of test statistics are lower as degrees of freedom increase. Subsequently, the research questions were answered comparing the exemplary dataset method to the power estimates found by programming the formulaic approach, as presented in the Optimal Design manual, but using the same degrees of freedom that the mixed model assumes in SAS. When the degrees of freedom were the same and no assumptions were violated, the power estimates obtained using the exemplary dataset method were nearly and practically equal the estimates obtained using the formulaic approach.

The choice of degrees of freedom method, therefore, can have a substantial impact on statistical power. In order to determine if any method is appropriate for use in

an analysis, researchers should first conduct power analyses where the effect size is set to zero. In other words, a "Type I error" condition should be implemented. If the percent of significant test statistics, which is now the error rate as opposed to power, exceeds the nominal value of alpha (.05), the method has an inflated error rate and, thus, an inflated power rate. How much the error rate may deviate from alpha in Type I error conditions has been the subject of some debate. Bradley (1978) proposed a "liberal criterion" of 0.5*alpha to 1.5*alpha (e.g. 2.5% to 7.5%), as the range within which error rates may be considered nominal. Serlin (2000) proposed a more conservative range of 3.5% to 6.5%, but Bradley's liberal criterion is more often cited.

In this study, a Type I error condition was conducted for the first cell of the design in which an empirical power analysis was carried out at the F-statistics for each replication were used to calculate *p*-values based on *J*-2 degrees of freedom as well as the containment (SAS's default), between-within (used in the empirical and exemplary portions of this study), and Satterthwaite methods. The Kenward-Roger and Satterthwaite methods were equivalent. While *J*-2 (alpha=4.74%) and Satterthwaite (alpha=5.02%) had error rates very near to 5%, the containment and between-within methods both had error rates of approximately 6.8%. While the between-within error rate were outside of Serlin's (2000) criterion, they were within Bradley's (1978) liberal criterion. However, confidence intervals around the error rates for containment and between-within did not contain .05. Another Type I error condition was then performed for the last cell of the design, with ICC = .15, 15% attrition per interval, unbalanced treatment groups, and skewed cluster sizes. In this condition, all methods had error rates within Bradley's liberal criterion as well as Serlin's criterion. The method used in Optimal Design, *J*-2,

showed a lower error rate than nominal in this condition (.043), indicating that the power rate is too conservative. The results found here provide some evidence that the Satterthwaite or Kenward-Roger adjustments are the most appropriate options for hierarchical linear models. Although power estimates may be inflated, the between-within method was used in this study in order to ensure the comparability of the empirical and exemplary dataset methods for power analysis. See Appendix D for a summary of results in the Type I error conditions investigated.

Comparison of the Empirical and Exemplary Dataset Methods

Both the exemplary dataset method and the empirical power analysis method were performed in the original 54 cells of the design, and the two methods were expected to provide equal estimates of power across all conditions. Figures 6 to 8, however, show the empirical power estimates to be slightly lower than the exemplary dataset estimates in many cases. The empirical estimates were below exemplary estimates in 41 of the 54 cells (76%). Empirical power estimates ranged from 1.6% below the exemplary dataset estimates to .1% above the exemplary estimates. To investigate the reason for the differences, the first cell of the design was investigated where the empirical estimate was 1.1% below the exemplary estimate (the star shown in the top left corner of Figure 6). In each cell the empirical analysis was carried out with 25,600 replications so that power estimates would be accurate when rounded to the nearest percent. Although a 1% difference may be practically insignificant, these small differences were observed in several cells of the design. In order to determine if the difference is due to sampling variation in the empirical power analyses, a second empirical analysis was performed for the first cell using 50,000 replications, which produced a power estimate that was 0.9%

below the exemplary estimate. The decreased difference as a result of increasing the number of replications in this cell along with the fact that most of the differences found between the two methods are quite small indicates that any differences will likely approach zero as the number of replications in the empirical analysis increases asymptotically. The empirical power analysis method and the exemplary dataset method may be assumed to provide equivalent estimates of power and sample size, as long as the number of replications in an empirical power analysis is sufficiently large. A different random seed was used in the second execution of the cell, which may have limited the convergence of the two power estimates.

Although the resulting estimates of the empirical and exemplary methods may be equal, the amount and type of work involved in the two methods are considerably different. The empirical method may take a large amount of time to complete its replications. Approximately seven hours were needed to carry out 25,600 replications using a computer with 8 GB of RAM (memory) and a 2.8 GHz processor. For such a large number of replications, seven hours is a reasonable amount of time, especially when the program can be run over night. However, if one small mistake is made in the syntax then the entire set of replications needs to be performed again, and several mistakes can greatly increase the time and computational burden. If twice the number of replications is performed, the program may take considerably longer than twice the amount of time, as the amount of available working memory in SAS gets smaller as the program runs longer. The exemplary dataset method has the advantage of providing results nearly instantaneously.

A potential drawback of the exemplary dataset method is that under very complex design conditions it can be difficult to properly set up the exemplary dataset. Under the condition where cluster sizes were normally distributed or negatively skewed, the distribution of cluster sizes, and the number of clusters at each cluster size in the treatment and control conditions, had to be manually programmed. A randomly generated set of cluster sizes is not acceptable because any difference in the distribution of cluster sizes across treatment conditions from exactly the same distribution of cluster sizes across treatment groups results in an unanticipated imbalance and the dataset is no longer exemplary. For example, when there are 25 clusters the analyst must place those 25 clusters in such a way that the treatment and control groups are as evenly divided as possible in terms of the resulting total sample size in each group. Under complex conditions such as these, the empirical power method may be considerably easier to implement because random generation of design conditions will average out over the full set of replications. Using the exemplary dataset method, there is only one replication and nothing can be randomly generated.

One reason why the empirical power method may be preferred from a theoretical perspective is that it more closely approximates reality in that no study will ever have perfect, “exemplary” data. Most samples and resulting data sets will be slightly different in some way from what was originally intended because of random assignment, outliers, data entry problems, or the logistics of carrying out a study in field settings. No data set will have exactly the fixed effects or variance components envisioned prior to performing a study. Empirical analyses better reflect these realities because the data set simulated in each replication is slightly different from the last, the next, and every replication in the

entire set. It is worth noting, however, that under the conditions in this study where assumptions were not violated, the exemplary dataset method provided power estimates that were approximately equal to those obtained using power formulae. The empirical estimates deviated slightly from the theoretically correct values of power. This may indicate that the exemplary dataset method provides power estimates that are exact without having to perform empirical analyses with hundreds of thousands of replications. The choice may simply come down to which method a particular researcher is more familiar and comfortable with, because the estimates obtained are essentially equal.

Another limitation of the exemplary dataset method is that approximate degree of freedom methods, including Satterthwaite and Kenward-Roger, cannot be accommodated using the MIXED procedure in SAS. If the degrees of freedom can be approximated manually, they can be used when power is calculated via the probability density function of the F-distribution. However, the approximate covariance matrix of the vector of fixed effects must be determined in order to do so. Although there is no direct evidence available in social science literature regarding the acceptability of these adjustments because they are most often implemented in small sample studies (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009), when designs are imbalanced in treatment conditions, cluster sizes, or the number of time points across individuals, these methods may provide much more appropriate values for the denominator degrees of freedom for testing fixed effects. Furthermore, the Type I error conditions presented in this discussion indicated that while error rates using the between-within or containment methods were inflated, the Satterthwaite and Kenward-Roger approximations preserved error rates to the nominal level. Many studies that plan analyses utilizing HLM use one of these

methods. A power analysis method that most accurately reflects the design and analysis conditions that will be implemented in a study should provide the most precise estimates of statistical power. Empirical power analysis methods in SAS can accommodate exact or approximate degrees of freedom.

The exemplary dataset method should only be implemented in situations where an empirical analysis, under the same experimental and analytic conditions, maintained nominal error rates in the Type I error condition. If an empirical power analysis using between-within degrees of freedom has acceptable levels of Type I error, then power values obtained using an exemplary dataset analysis under the same conditions may be considered valid. The final, but perhaps most important, drawback of the exemplary dataset method is that it does not facilitate Type I error investigations. Because the fixed effects are precisely zero when the effect size is set to zero, there can be no deviation in Type I error from the nominal rate. An empirical analysis that matches the design conditions of interest is the only way to verify that the power values obtained under the exemplary dataset analysis are not inflated. Thus, empirical analyses are the standard against which all other power analysis methods should be compared.

General Discussion

The current study has demonstrated the importance of identifying the design conditions of a planned study and evaluating whether or not those conditions are accurately reflected in the determination of power and sample size. Researchers intending to calculate power for a longitudinal cluster randomized trial, as long as treatment groups are balanced and no attrition is expected, software programs such as Optimal Design will likely provide accurate estimates of power and sample size. What is important for users

to be aware of is that formulaic power analysis methods make simplifying assumptions which are often violated in practice. The effects of violations on power may not always affect power, as was found here under conditions of unequal cluster sizes, but power is adversely affected by common occurrences like unequal treatment group sizes and especially attrition. The effect of unequal cluster sizes could play a more significant role if the distribution of cluster sizes is not evenly spread across treatment conditions. Failing to account for attrition in any longitudinal study will have a definite impact on power. The magnitude of that impact will vary depending upon the length of the study, with lower attrition rates expected for shorter studies, and depending upon the nature of the study. Research studies that require more of a participant's time for measurement, have more measurement occasions, or that deal with sensitive issues may experience higher attrition rates. The magnitude of the impact of attrition on statistical power was found in this study to also depend on other design considerations, including treatment group balance, cluster size equality, and the amount of variability attributable to clusters. This suggests that ICC is not the only important moderator of the bias imposed as a result of failing to account for these design considerations. Each element of the design may have an impact on the bias imposed by violating assumptions made concerning other elements. No known heuristic -- such as oversampling by the amount of expected attrition or the amount of lost information -- can accurately reflect the increased sample size needed to account for attrition without taking into account the full context of a particular study.

There are many more possible design considerations than what were considered in this study that may have an effect on power. Formulaic power methods as they are currently operationalized in software such as Optimal Design assume equal intercept and

slope variances across treatment groups and across clusters. Formulaic methods assume intervals are equally spaced during the course of a study and that intervals are precisely equal across all individuals. Although formulas are available or can be derived that account for these conditions, they are not implemented in any one comprehensive program. Empirical and exemplary dataset power methods can accommodate all of these design considerations. In some cases, the conditions may be somewhat difficult to implement, but any design element possible can in some way be accounted for using the empirical or exemplary dataset methods.

An educational researcher performing a power analysis needs to take a critical look at all aspects of their research design. Funding agencies put a great deal of resources in the hands of researchers and analysts. When foreseeable design considerations are not carefully considered, the estimated sample size may be too low to find an effect that truly exists in the population. The result is often a waste of valuable resources. Funding agencies and research institutions also place a great deal of faith on researchers that they are not putting too many human participants at risk due to the seemingly inconsequential act of overpowering a study.

References

Atkins, D. C. (2009) Clinical trials methodology: Randomization, intent-to-treat, and random-effects regression. *Depression and Anxiety*, *26*, 697-700.

Biostat Inc. (2001). *Power And Precision™ software manual*. Englewood, NJ: Biostat Inc.

Bird, K. D, & Hall, W. (1986). Statistical power in psychiatric research. *Australian and New Zealand Journal of Psychiatry*, *20*, 189-200.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Burke, C. J. (1953). A brief note on one-tailed tests. *Psychological Bulletin*, *50*, 384-387.

Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: An introduction to the latent growth curve model. *Understanding Statistics*, *2*, 177-203.

Candel, M. J. J. M, Van Breukelen, G. J. P., Kotova, L., & Berger, M. P. F. (2008). Optimality of unequal cluster sizes in multilevel studies with small sample sizes. *Communications in Statistics: Simulation and Computation*, *37,* 222-239.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cowles, M. (2001). *Statistics in psychology: An historical perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553-558.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"-or should we? *Psychological Bulletin*, *74*, 68-80.

Curran, P. J., & Muthén, B. O. (1999). The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology*, *27*, 567-595.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*, 159-165.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., . . . Zill, N. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*, 558- 580.

Elashoff, J. D. (2007). nQuery Advisor (Version 7) [Computer software]. Cork, Ireland: Statistical Solutions.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1-11.

Faul, E., Erdfelder, F., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372-384.

Flannery, D. J., Vazsonyi, A. T., Liau, A. K., Guo, S., Powell, K. E., Atha, H., . . . Embry, D. (2003). Initial behavior outcomes for the PeaceBuilders universal

school-based violence prevention program. *Developmental Psychology*, *39*, 292–308.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499-510.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Hammarberg, A., & Hagekull, B. (2002). The relation between pre-school teachers' classroom experiences and their perceived control over child behaviour. *Early Child Development and Care*, *172*, 625–634.

Harris, R. J. (1975). *A primer on multivariate statistics.* New York, NY: Academic.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, *24*, 70-93.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61-85.

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*, 639–657.

Hevey, D., & McGee, H. M. (1998). The effect size statistic: Useful in health outcomes research? *Journal of Health Psychology*, *3*, 163-170.

Hill, N. E., Castellino, D. R., Lansford, J. E., Nowlin, P., Dodge, K. A., Bates, J. E., & Pettit, G. S. (2004). Parent academic involvement as related to school behavior, achievement, and aspirations: Demographic variations across adolescence. *Child Development*, *75*, 1491–1509.

Hintze, J. (2008). PASS 2008 [Computer software]. Kaysville, UT: NCSS, LLC. Retrieved from http://www.ncss.com

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

Houle, T. T., Penzien, D. B., & Houle, C. K (2005) Statistical power and sample size estimation for headache research: An overview and power calculation tools. *Headache*, *45*, 414-418.

Institute of Education Sciences. (2011). *Resources for researchers: Methodological resources*. Retrieved from http://ies.ed.gov/funding/resources.asp

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, *2*, 178–193.

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*, 755-766.

Kenny, D. A. (1996). The design and analysis of social-interaction research. *Annual Review of Psychology*, *47*, 59-86.

Kenny, D. A., Kashy, D., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 233-265). New York, NY: McGraw-Hill.

Kimmell, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*, 351-353.

Kraemer, H. C. (1985). A strategy to teach the concept and application of power of statistical tests. *Journal of Educational Statistics*, *10*, 173-195.

Koehler, E., Brown, E., & Haneuse, S. J. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*, 155-162.

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education*, *78*, 291-317.

Laczo, R. M., Sackett, P. R., Bobko, P., & Cortina, J. M. (2005). A comment on sampling error in the standardized mean difference with unequal sample sizes: Avoiding potential errors in meta-analytic and primary research. *Journal of Applied Psychology*, *90*, 758–764.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193.

Leidy, N. K, & Weissfeld, L. A. (1991). Sample sizes and power computation for clinical intervention trials. *Western Journal of Nursing Research*, *13*, 138-144.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Loehlin, J. C. (2004). *Latent variable models* (4th ed., pp. 70-73). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Longford, N. (1993). *Random coefficient models*. Oxford, England: Oxford University Press.

Lunneborg C. E., & Tousignant, J. P. (1985). Efron's bootstrap with application to the repeated measures design. *Multivariate Behavioral Research*, *20*, 161-178.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149.

MacCallum, R. C. & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research*, *32*, 193-210.

Malmgren, K. W., & Gagnon, J. C. (2005). School mobility and students with emotional disturbance. *Journal of Child and Family Studies*, *14*, 299-312.

Man-Son-Hing, M., Laupacis, A., O'Rourke, K., Molnar, F. J., Mahon, J., Chan, K. B. Y., & Wells, G. (2002). Determination of the clinical importance of study results: A review. *Journal of General Internal Medicine*, *17*, 469–476.

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, *58*, 575-585.

Mead, R. (1988). *The design of experiments: Statistical principals for practical applications*. (p. 24). New York, NY: Cambridge University Press.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*, 611–635.

Melby, J. N., & Conger, R. D. (1996). Parental behaviors and adolescent academic performance: A longitudinal analysis. *Journal of Research on Adolescence*, *6*, 113–137.

Moerbeek, M., Van Breukelen, G. J. P., Berger, M. P. F., & Ausems, M. (2003). Optimal sample sizes in experimental designs with individuals nested within clusters. *Understanding Statistics*, *2*, 151-175.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

Murphy, K. R., & Myors, B. (2004*). Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd Ed.). Mahwah, NJ: Erlbaum.

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.

Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, *7*, 103-108.

National Institutes of Health. (2011). *Lifestyle interventions in overweight and obese pregnant women consortium (U01)*. Retrieved from http://grants.nih.gov/grants/guide/rfa-files/RFA-DK-10-014.html

Nye, B., Hedges, V. E., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, *37*, 123-151.

O'Brien, R. G. (1998*). A tour of UnifyPow: A SAS module/macro for sample-size analysis*. Proceedings of the 23rd SAS Users Group International Conference, Cary, NC, SAS Institute, 1346-1355.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for t tests through multivariate hypotheses. In L.K. Edwards (Ed.), *Applied Analysis of Variance in Behavioral Science* (pp. 297-344). New York, NY: Marcel Dekker.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447.

Overall, J. E., Shobaki, G., Shivakumar, C., & Steel, J. (1998). Adjusting sample size for anticipated dropouts in clinical trials. *Psychopharmacology Bulletin*, *34*, 25-33.

Prinz, R. J., & Miller, G. E. (1994). Family-based treatment for childhood antisocial behavior: Experimental influences on dropout and engagement. *Journal of Consulting and Clinical Psychology*, *62*, 645-650.

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

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

Raudenbush, S. W., Bryk, A. S, & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199-213.

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387-401.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Champion, K. M., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, *23*, 10–26.

Rigby, A. S., & Vail, A. (1998). Statistical methods in epidemiology. II: A commonsense approach to sample size estimation. *Disability and Rehabilitation*, *20*, 405-410.

Robey, R. R. (2004). Reporting point and interval estimates of effect-size for planned contrasts: fixed within effect analyses of variance. *Journal of Fluency Disorders*, *29*, 307-341

SAS Institute Inc. (2010). *OnlineDoc® 9.2*. Cary, NC: SAS Institute Inc.

Saunders, C. L., Bishop, D. T., & Barrett, J. H. (2003). Sample size calculations for main effects and interactions in case–control studies using Stata's nchi2 and npnchi2 functions. *The Stata Journal*, *3*, 47–56.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2001). Approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. *Proceedings of the 26th SAS Users Group International Conference*, *Cary*, *NC*, *SAS Institute*, *Paper 262-26*.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, *5*, 230-240.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323-355.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). *Optimal Design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. Available from http://www.wtgrantfoundation.org/resources/overview/research_tools/research_tools

Stroup, W. W. (1999). Mixed model procedures to assess power, precision, and sample size in the design of experiments. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 15–24.

Stroup, W. W. (2002). Power analysis based on spatial mixed effects analysis: a tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 491–511.

Theommes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling*, *17*, 510-534.

van Doesum, K. T. M., Riksen-Walraven, J. M., Hosman, C. M. H., & Hoefnagels, C. (2008). A randomized controlled trial of a home-visiting intervention aimed at

preventing relationship problems in depressed mothers and their infants. *Child Development*, *79*, 547–561.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, *39*, 95-101.

Zelin, A., & Stubbs, R., (2005). Cluster sampling: A false economy? *International Journal of Market Research*, *47*, 503-524.

APPENDIX A: SYNTAX FOR EMPIRICAL POWER ANALYSIS

```
%macro empirical;
%let reps=25600;        *how many replications-need 25600;

%let set_n2=16;         *how many clusters;
%let set_icc=0.05;      *how much variance in slopes at the cluster
level? .05, .10, or .15;
%let set_prop_treat=.5; *proportion of clusters in treatment condition
      .5 or .7;
%let set_prop_missing=0; *Proportion of droput per interval use 0, .05,
or .15;
%let set_cluster_condition=1;       *1: equal at 20
                                     2: normal(20,2)
                                     3: negative skew;

data results_condition11;input rep Pintvar Psvar Cintvar Csvar residual
sICC DF parameter pval;run;
%do r=1 %to &reps;
proc iml;
n1=20;                  *how large are the clusters;
sigma=.5;               *residual variance;
es=.5;                  *effect size;

n2=&set_n2.;            *create iml constants to make syntax more
concise;
icc=&set_icc.;
prop_treat =&set_prop_treat.;
prop_missing=&set_prop_missing.;
cluster_condition=&set_cluster_condition.;

csizes=J(n2,1);

if cluster_condition=1 then csizes[,1]=n1;      *CLUSTER SIZE CONDITION
1: equal cluster sizes;
if cluster_condition=2 then do;                 *CLUSTER SIZE CONDITION
2: normal(20,2) cluster sizes;
      do j=1 to n2;
            csizes[j,1]=round(n1+2*rannor(0));
            if csizes[j,1]<15 then csizes[j,1]=15; *no smaller than 15;
            if csizes[j,1]>25 then csizes[j,1]=25; *no larger than 25;
      end;
end;
if cluster_condition=3 then do;                 *CLUSTER SIZE CONDITION
3: negatively skewed cluster sizes;
      toggle=2;
      do j=1 to n2;
            csizes[j,1]=round(n1+2*rannor(0));
            if csizes[j,1]>20 then do;
            toggle=1/toggle;
            if toggle=2 then csizes[j,1]=round(20-abs(20-csizes[j,1]));

            *flop to other side;
            if toggle=.5 then csizes[j,1]=round(csizes[j,1]-.5*abs(20-
            csizes[j,1]));
            *half the distance to the mean;
```

```
            end;
            if csizes[j,1]<15 then csizes[j,1]=15; *no smaller than 15;
            if csizes[j,1]>22 then csizes[j,1]=22; *no larger than 22;
      end;
end;

      N=sum(csizes[,1]);            *calculate total sample size;
      control_Cs=round(n2*(1-prop_treat));
      treat_Cs=n2-control_Cs;       *how many clusters in each group;
      S=J(4*N,5);                   *set up data matrix shell;
      Cvar=J(n2,2);*matrix for cluster intercept and slope variances;
      Pvar=J(N,2); *matrix for person intercept and slope variances;
      index=0;id=0;
      groups=J(1,2);groups[,]=0;

if (prop_missing>0) then do;
      missing=J(N,3);
      do m=1 to N;
            missing[m,1]=m;                *person id;
            missing[m,2]=ranuni(0);        *random draw;

      end;
      missing[,3]=rank(missing[,2]);      *rank order by random draws;
end;

do j=1 to n2;
      group=0;if (j>control_Cs) then group=1;
      Cvar[j,1]=sqrt(.1)*rannor(0); *cluster-level intercept variance;
      Cvar[j,2]=sqrt(icc)*rannor(0); *cluster-level slope variance;

      do i=1 to csizes[j,1];
            id=id+1;
            Pvar[id,1]=sqrt(.2)*rannor(0);            *person-level
intercept variance;
            Pvar[id,2]=sqrt((1-icc))*rannor(0);       *person-level
slope variance;
            do t=1 to 4;
                  index=index+1;
                  S[index,1]=id;
                  S[index,2]=j;
                  S[index,3]=t-1;
                  S[index,4]=group;
      S[index,5]=(Cvar[j,2]+Pvar[id,2])*(t-1)+(es)*group*(t-
1)+sqrt(.5)*rannor(0)+Cvar[j,1]+Pvar[id,1];
                  if (prop_missing>0) then do;
                        if (t>1) then do;
                              if
(missing[id,3]<=round((prop_missing*(t-1)*N))) then S[index,5]=.;
                        end;
                  end;
            end;
      end;
end;
CREATE simdata from S[COLNAME={id cluster time group y}];APPEND FROM S;
quit; *end iml portion for this replication;

      PROC MIXED data=simdata noclprint covtest;      *run the model;
```

```
      CLASS CLUSTER id;
      MODEL Y= time group time*group/solution ddfm=bw;
      RANDOM int time/sub=id(cluster) TYPE=VC;
      RANDOM int time/sub=cluster TYPE=VC;
      ODS OUTPUT CovParms=rand;           *save variance components;
      ODS OUTPUT solutionf=fixed;         *save fixed effects;
      ODS OUTPUT tests3=fixedF;
      RUN;

      data Rand;set Rand;keep Estimate;run;proc transpose data=rand
out=rand;run;data rand;set rand;rep=&r;
      Pintvar=COL1;Psvar=COL2;Cintvar=COL3;Csvar=COL4;residual=COL5;
      sICC=Csvar/(Csvar+Psvar);keep rep Pintvar Psvar Cintvar Csvar
residual sICC;run;
      data fixed;set fixed;if
Effect='TIME*GROUP';parameter=Estimate;pval=probt;keep DF parameter
pval;run;
      data res;merge rand fixed;run;
      proc append base=results_condition11;run;
      DM 'CLEAR LOG';
      DM 'CLEAR OUTPUT';
%end; *end of loop for this replication;

data results_condition11;set results_condition11;sig=0;if pval<=.05
then sig=1; *for calculating power;
 n2=&set_n2.;     *pass conditions on to results file just for keeping
track;
 icc=&set_icc.;
 prop_treat=&set_prop_treat.;
 prop_missing=&set_prop_missing.;
 cluster_condition=&set_cluster_condition.;
run;
%mend empirical;
%empirical;
```

APPENDIX B: SYNTAX FOR EXEMPLARY DATASET POWER ANALYSIS

```
%macro exemplary;
%let mat_prop_treat=.5 .7;          *proportion of clusters in
treatment condition;
%let mat_prop_missing=0 0.05 0.15;  *Proportion of droput per interval;
%let mat_cluster_condition=1 2 3;
%let cond=6;

data Results_icc05;input effect $20. numdf dendf fvalue probf alpha
ncparm fcrit
power_exemplary condition n2 icc prop_missing prop_treat
cluster_condition;run;

%do a=1 %to 3;                      *attrition loop;
%let cond=%eval(&cond-6+10);
%do pt=1 %to 2;                     *treatment balance loop;
%do cs=1 %to 3;                     *cluster condition loop;
%let cond=%eval(&cond+1);
%let att = %scan(&mat_prop_missing,&a,%str( ));
%let ptr = %scan(&mat_prop_treat,&pt,%str( ));
%let csz = %scan(&mat_cluster_condition,&cs,%str( ));
proc iml;

n2=13;                  *cluster size;
icc=.05;                *intraclass correlation;
n1=20;                  *how large are the clusters;
sigma=.5;               *residual variance;
es=.5;                  *effect size;
cond=&cond.;
prop_missing=&att.;
prop_treat=&ptr.;
cluster_condition=&csz.;

parms=cond||n2||icc||prop_missing||prop_treat||cluster_condition;

csizes=J(n2,1);
if cluster_condition=1 then csizes[,1]=n1;      *CLUSTER SIZE CONDITION
1: equal cluster sizes;
if cluster_condition=2 then do;                 *CLUSTER SIZE CONDITION
2: normal(20,2) cluster sizes;
normsizes={0 0 1 2 2 3 2 2 1 0 0,0 0 1 2 2 4 2 2 1 0 0,
0 0 1 2 2 4 3 2 1 0 0,0 0 1 2 3 4 3 2 1 0 0,0 0 1 2 3 5 3 2 1 0 0,
0 1 1 2 3 4 3 2 1 1 0,0 1 1 2 3 5 3 2 1 1 0,0 1 1 2 4 4 4 2 1 1 0,
0 1 1 2 4 5 4 2 1 1 0,0 1 2 2 4 4 4 2 2 1 0,0 1 2 2 4 5 4 2 2 1 0,
0 1 2 3 4 4 4 3 2 1 0,1 1 2 2 4 5 4 2 2 1 1,1 1 2 2 4 6 4 2 2 1 1,
1 1 2 2 5 5 5 2 2 1 1,1 1 2 3 4 6 4 3 2 1 1,1 1 2 3 5 5 5 3 2 1 1,
1 1 2 3 5 6 5 3 2 1 1,1 1 2 4 5 5 5 4 2 1 1,1 1 2 4 5 6 5 4 2 1 1,
1 1 2 4 5 7 5 4 2 1 1,1 1 2 4 6 6 6 4 2 1 1,1 1 2 4 6 7 6 4 2 1 1,
1 1 3 4 6 6 6 4 3 1 1,1 1 3 4 6 7 6 4 3 1 1,1 2 3 4 6 6 6 4 3 2 1,
1 2 3 4 6 7 6 4 3 2 1,1 2 3 5 6 6 6 5 3 2 1,1 2 3 5 6 7 6 5 3 2 1};
normsize=normsizes[n2-12,];
index=0;
toggle=2;                       *alternate treatment and control;
do s=15 to 25;
do j=1 to normsize[s-14];
```

```
      if toggle=2 then do;index=index+1;csizes[index,1]=s;end;
      if toggle=.5 then do;csizes[index+round(n2/2),1]=s;end;
      toggle=1/toggle;
end;end;
end;

if cluster_condition=3 then do;                 *CLUSTER SIZE CONDITION
3: negatively skewed cluster sizes;
skewsizes={0 1 1 2 2 3 2 2,0 1 1 2 3 3 2 2,0 1 1 2 3 3 3 2,
0 1 1 2 3 4 3 2,0 1 1 2 4 4 3 2,0 1 1 2 4 4 4 2,0 1 1 2 4 5 4 2,
0 1 1 2 5 5 4 2,0 1 1 2 5 5 5 2,0 1 2 3 5 5 4 2,1 1 2 3 5 5 4 2,
1 1 2 3 5 5 5 2,1 1 2 3 5 6 5 2,1 1 2 4 5 6 5 2,1 1 2 4 5 6 5 3,
1 1 3 4 5 6 5 3,1 2 3 4 5 6 5 3,1 2 3 4 6 6 5 3,1 2 3 4 6 7 5 3,
1 2 3 4 6 7 5 4,1 2 3 4 6 7 6 4,1 2 3 4 7 7 6 4,1 2 3 4 7 7 7 4,
1 2 3 5 7 7 7 4,1 2 3 5 7 8 7 4,1 2 3 5 8 8 7 4,1 2 3 5 8 8 8 4,
1 2 4 5 8 8 8 4,1 2 4 6 8 9 7 4};
skewsize=skewsizes[n2-12,];
index=0;
toggle=2;
do s=15 to 22;
do j=1 to skewsize[s-14];
      if toggle=2 then do;index=index+1;csizes[index,1]=s;end;
      if toggle=.5 then do;csizes[index+round(n2/2),1]=s;end;
      toggle=1/toggle;
end;end;
end;

N=sum(csizes[,1]);                  *calculate total sample size;
control_Cs=round(n2*(1-prop_treat));
treat_Cs=n2-control_Cs;             *how many clusters in each group;

if (prop_missing>0) then do;
      missing=J(N,3);
      do m=1 to N;
            missing[m,1]=m;                     *person id;
            missing[m,2]=ranuni(0);             *random draw;
      end;
      missing[,3]=rank(missing[,2]);            *rank by random draws;
end;

Se=J(4*N,5);                        *set up data matrix shell;
index=0;id=0;
do j=1 to n2;group=0;if (j>control_Cs) then group=1;
      do i=1 to csizes[j,1];
            id=id+1;
            do t=1 to 4;
                  index=index+1;
                  Se[index,1]=id;
                  Se[index,2]=j;
                  Se[index,3]=t-1;
                  Se[index,4]=group;
                  Se[index,5]=(es)*group*(t-1);
                  if (prop_missing>0) then do;
                        if (t>1) then do;
                        if (missing[id,3]<=round((prop_missing*(t-
1)*N))) then Se[index,5]=.;end;
                        end;
```

```
            end;
      end;
end;
CREATE exempdata from Se[COLNAME={id cluster time group y}];APPEND FROM
Se;
CREATE parm from parms [COLNAME={condition n2 icc prop_missing
prop_treat cluster_condition}];APPEND from parms;close parm;
quit;

PROC MIXED data=exempdata noclprint;
CLASS CLUSTER id;
MODEL Y= time group time*group/solution ddfm=bw;
RANDOM int time/sub=id(cluster) TYPE=VC;
RANDOM int time/sub=cluster TYPE=VC;
PARMS (.2) (.95) (.1) (.05) (.5) / hold=1,2,3,4,5;    *Set the variance
components to match icc;
ODS OUTPUT tests3=Ftests;
RUN;

data Ftests; set Ftests;if Effect='TIME*GROUP';run;
data Ftests; set Ftests;
 alpha=0.05;
 ncparm=numdf*fvalue;                      *noncentrality parameter;
 fcrit=finv(1-alpha,numdf,dendf,0);        *critical F value;
 power_exemplary=1-probf(fcrit,numdf,dendf,ncparm);   *calculate power;
data Ftests;merge Ftests parm;run;

proc append base=results_icc05;run;
%end;%end;%end;
%mend exemplary;
%exemplary;
```

APPENDIX C. SIMULATION DIAGNOSTICS

Table C.1. Person Intercept Variance (Population Value)

		Treatment Balance					
		50/50			70/30		
		Cluster Balance			Cluster Balance		
ICC	Percent Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
.05	0%	0.1997 (0.2)	0.2001 (0.2)	0.1999 (0.2)	0.2006 (0.2)	0.1999 (0.2)	0.1999 (0.2)
	5%	0.2004 (0.2)	0.2000 (0.2)	0.2004 (0.2)	0.2001 (0.2)	0.1999 (0.2)	0.1996 (0.2)
	15%	0.2003 (0.2)	0.1997 (0.2)	0.1995 (0.2)	0.2005 (0.2)	0.1999 (0.2)	0.2001 (0.2)
.10	0%	0.2000 (0.2)	0.2001 (0.2)	0.1997 (0.2)	0.1997 (0.2)	0.1998 (0.2)	0.1994 (0.2)
	5%	0.2001 (0.2)	0.2001 (0.2)	0.2002 (0.2)	0.1998 (0.2)	0.2000 (0.2)	0.1995 (0.2)
	15%	0.1999 (0.2)	0.1999 (0.2)	0.1995 (0.2)	0.2001 (0.2)	0.2002 (0.2)	0.2005 (0.2)
.15	0%	0.1999 (0.2)	0.2001 (0.2)	0.1997 (0.2)	0.2001 (0.2)	0.2003 (0.2)	0.2002 (0.2)
	5%	0.1999 (0.2)	0.1998 (0.2)	0.1997 (0.2)	0.1999 (0.2)	0.1998 (0.2)	0.2000 (0.2)
	15%	0.2001 (0.2)	0.2000 (0.2)	0.2004 (0.2)	0.1998 (0.2)	0.1996 (0.2)	0.1999 (0.2)

Table C.2. Cluster Intercept Variance (Population Value)

		Treatment Balance					
		50/50			70/30		
		Cluster Balance			Cluster Balance		
ICC	Percent Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
.05	0%	0.0996 (0.1)	0.1000 (0.1)	0.0993 (0.1)	0.0998 (0.1)	0.1004 (0.1)	0.1002 (0.1)
	5%	0.1003 (0.1)	0.0997 (0.1)	0.0996 (0.1)	0.0998 (0.1)	0.1004 (0.1)	0.0998 (0.1)
	15%	0.1003 (0.1)	0.0996 (0.1)	0.0994 (0.1)	0.0999 (0.1)	0.1003 (0.1)	0.0996 (0.1)
.10	0%	0.1001 (0.1)	0.0995 (0.1)	0.0994 (0.1)	0.1003 (0.1)	0.1000 (0.1)	0.1003 (0.1)
	5%	0.1001 (0.1)	0.1000 (0.1)	0.1000 (0.1)	0.1005 (0.1)	0.1001 (0.1)	0.1003 (0.1)
	15%	0.0997 (0.1)	0.1003 (0.1)	0.0995 (0.1)	0.1003 (0.1)	0.0999 (0.1)	0.0998 (0.1)
.15	0%	0.1003 (0.1)	0.0997 (0.1)	0.1003 (0.1)	0.1001 (0.1)	0.1000 (0.1)	0.1000 (0.1)
	5%	0.0999 (0.1)	0.0999 (0.1)	0.0997 (0.1)	0.1000 (0.1)	0.1000 (0.1)	0.0999 (0.1)
	15%	0.1001 (0.1)	0.1000 (0.1)	0.1000 (0.1)	0.1002 (0.1)	0.0998 (0.1)	0.0999 (0.1)

Table C.3. Person Slope Variance (Population Value)

		Treatment Balance					
		50/50			70/30		
	Percent	Cluster Balance			Cluster Balance		
ICC	Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	0.9493 (0.95)	0.9492 (0.95)	0.9487 (0.95)	0.9496 (0.95)	0.9493 (0.95)	0.9493 (0.95)
.05	5%	0.9501 (0.95)	0.9502 (0.95)	0.9499 (0.95)	0.9493 (0.95)	0.9500 (0.95)	0.9484 (0.95)
	15%	0.9483 (0.95)	0.9484 (0.95)	0.9486 (0.95)	0.9481 (0.95)	0.9479 (0.95)	0.9483 (0.95)
	0%	0.8994 (0.9)	0.9002 (0.9)	0.9007 (0.9)	0.8999 (0.9)	0.8999 (0.9)	0.8999 (0.9)
.10	5%	0.9000 (0.9)	0.9012 (0.9)	0.8998 (0.9)	0.8997 (0.9)	0.9006 (0.9)	0.8998 (0.9)
	15%	0.8996 (0.9)	0.8992 (0.9)	0.9005 (0.9)	0.9004 (0.9)	0.9003 (0.9)	0.8999 (0.9)
	0%	0.8499 (0.85)	0.8502 (0.85)	0.8501 (0.85)	0.8498 (0.85)	0.8507 (0.85)	0.8499 (0.85)
.15	5%	0.8496 (0.85)	0.8497 (0.85)	0.8500 (0.85)	0.8502 (0.85)	0.8501 (0.85)	0.8500 (0.85)
	15%	0.8504 (0.85)	0.8495 (0.85)	0.8501 (0.85)	0.8505 (0.85)	0.8506 (0.85)	0.8501 (0.85)

Table C.4. Cluster Slope Variance (Population Value)

		Treatment Balance					
		50/50			70/30		
	Percent	Cluster Balance			Cluster Balance		
ICC	Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	0.0506 (0.05)	0.0510 (0.05)	0.0507 (0.05)	0.0511 (0.05)	0.0508 (0.05)	0.0508 (0.05)
.05	5%	0.0508 (0.05)	0.0503 (0.05)	0.0512 (0.05)	0.0511 (0.05)	0.0505 (0.05)	0.0515 (0.05)
	15%	0.0518 (0.05)	0.0517 (0.05)	0.0519 (0.05)	0.0519 (0.05)	0.0512 (0.05)	0.0521 (0.05)
	0%	0.1001 (0.1)	0.0996 (0.1)	0.0996 (0.1)	0.1003 (0.1)	0.0998 (0.1)	0.1002 (0.1)
.10	5%	0.1001 (0.1)	0.0995 (0.1)	0.1005 (0.1)	0.1000 (0.1)	0.0996 (0.1)	0.0999 (0.1)
	15%	0.1002 (0.1)	0.1003 (0.1)	0.0999 (0.1)	0.1000 (0.1)	0.1000 (0.1)	0.1003 (0.1)
	0%	0.1498 (0.15)	0.1494 (0.15)	0.1499 (0.15)	0.1502 (0.15)	0.1498 (0.15)	0.1502 (0.15)
.15	5%	0.1499 (0.15)	0.1498 (0.15)	0.1501 (0.15)	0.1501 (0.15)	0.1498 (0.15)	0.1496 (0.15)
	15%	0.1498 (0.15)	0.1494 (0.15)	0.1493 (0.15)	0.1498 (0.15)	0.1501 (0.15)	0.1500 (0.15)

Table C.5. Intraclass Correlation Coefficient (Population Value)

		Treatment Balance					
		50/50			70/30		
	Percent	Cluster Balance			Cluster Balance		
ICC	Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	0.0499 (0.05)	0.0502 (0.05)	0.0500 (0.05)	0.0503 (0.05)	0.0500 (0.05)	0.0501 (0.05)
.05	5%	0.0500 (0.05)	0.0495 (0.05)	0.0503 (0.05)	0.0503 (0.05)	0.0497 (0.05)	0.0507 (0.05)
	15%	0.0510 (0.05)	0.0508 (0.05)	0.0510 (0.05)	0.0510 (0.05)	0.0503 (0.05)	0.0512 (0.05)
	0%	0.0987 (0.1)	0.0982 (0.1)	0.0981 (0.1)	0.0989 (0.1)	0.0985 (0.1)	0.0988 (0.1)
.10	5%	0.0986 (0.1)	0.0979 (0.1)	0.0990 (0.1)	0.0986 (0.1)	0.0981 (0.1)	0.0985 (0.1)
	15%	0.0987 (0.1)	0.0989 (0.1)	0.0983 (0.1)	0.0985 (0.1)	0.0985 (0.1)	0.0988 (0.1)
	0%	0.1479 (0.15)	0.1476 (0.15)	0.1480 (0.15)	0.1483 (0.15)	0.1478 (0.15)	0.1482 (0.15)
.15	5%	0.1481 (0.15)	0.1479 (0.15)	0.1481 (0.15)	0.1480 (0.15)	0.1479 (0.15)	0.1477 (0.15)
	15%	0.1478 (0.15)	0.1475 (0.15)	0.1473 (0.15)	0.1477 (0.15)	0.1479 (0.15)	0.1479 (0.15)

Table C.6. Time*Group Interaction (Population Value)

		Treatment Balance					
		50/50			70/30		
	Percent	Cluster Balance			Cluster Balance		
ICC	Attrition	Equal	Normal	Skewed	Equal	Normal	Skewed
	0%	0.4987 (0.5)	0.4989 (0.5)	0.4989 (0.5)	0.5009 (0.5)	0.4989 (0.5)	0.4991 (0.5)
.05	5%	0.5003 (0.5)	0.4997 (0.5)	0.5015 (0.5)	0.5015 (0.5)	0.5005 (0.5)	0.4997 (0.5)
	15%	0.5004 (0.5)	0.5007 (0.5)	0.4993 (0.5)	0.4983 (0.5)	0.4989 (0.5)	0.4993 (0.5)
	0%	0.4988 (0.5)	0.5003 (0.5)	0.4988 (0.5)	0.4986 (0.5)	0.5001 (0.5)	0.4996 (0.5)
.10	5%	0.5005 (0.5)	0.5022 (0.5)	0.4998 (0.5)	0.5006 (0.5)	0.5008 (0.5)	0.4993 (0.5)
	15%	0.5003 (0.5)	0.5016 (0.5)	0.4985 (0.5)	0.5007 (0.5)	0.4973 (0.5)	0.5005 (0.5)
	0%	0.4989 (0.5)	0.5006 (0.5)	0.5003 (0.5)	0.5004 (0.5)	0.5016 (0.5)	0.5021 (0.5)
.15	5%	0.5012 (0.5)	0.4995 (0.5)	0.4998 (0.5)	0.4997 (0.5)	0.4990 (0.5)	0.4999 (0.5)
	15%	0.5000 (0.5)	0.5000 (0.5)	0.5004 (0.5)	0.4984 (0.5)	0.5000 (0.5)	0.4988 (0.5)

APPENDIX D. TYPE I ERROR CONDITION SUMMARY

Table D.1. Type I error rates under ICC of .05, no attrition, balanced treatment groups, and equal cluster sizes.

DF	Error	.05-Error	2*SE CI Limits		Bradley's Liberal Criterion	Power
HLM_OD (J-2)	0.04736	0.0026	(.04467, .05004)	ok	ok	0.82416
Between-Within	0.06839	-0.0184	(.06520, .07158)	inflated	ok	0.86712
Containment	0.06811	-0.0181	(.06492, .07129)	inflated	ok	0.86676
Satterthwaite	0.05021	-0.0002	(.04744, .05279)	ok	ok	0.82544
Kenward-Roger	0.05021	-0.0002	(.04744, .05279)	ok	ok	0.82544

Table D.2. Type I error rates under ICC of .15, 15% attrition per interval, unbalanced treatment groups, and negatively skewed cluster sizes.

DF	Error	.05-Error	2*SE CI Limits		Bradley's Liberal Criterion	Power
HLM_OD (J-2)	0.04279	0.0072	(.04023, .04535)	too small	ok	0.67128
Between-Within	0.06279	-0.0128	(.05972, .06586)	inflated	ok	0.73280
Containment	0.06248	-0.0125	(.05942, .06555)	inflated	ok	0.73196
Satterthwaite	0.05170	-0.0017	(.04890, .05450)	ok	ok	0.70276
Kenward-Roger	0.05158	-0.0016	(.04878, .05437)	ok	ok	0.70276