5-1-2011

# Comparison of multivariate methods for measuring change from pretest to posttest

Justin Leslie Rogers
*University of Northern Colorado*

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

# A COMPARISON OF MULTIVARIATE METHODS FOR MEASURING CHANGE FROM PRETEST TO POSTTEST

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Justin Leslie Rogers

College of Education and Behavioral Sciences
School of Educational Research, Leadership and Technology
Program of Applied Statistics and Research Methods

May, 2011

This Dissertation by: Justin Leslie Rogers

Entitled: *A Comparison Of Multivariate Methods For Measuring Change From Pretest To Posttest*

has been approved as meeting the requirements for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences in School of Educational Research, Leadership and Technology, Program of Applied Statistics and Research Methods

Accepted by the Doctoral Committee:

_____
Daniel Mundfrom, Ph.D., Chair

_____
Jamis Perrett, Ph.D., Committee Member

_____
Jay Schaffer, Ph.D., Committee Member

_____
Robert Heiny, Ph.D., Faculty Representative

Date of Dissertation Defense March 31, 2011

Accepted by the Graduate School

_____
Robbyn R. Wacker, Ph.D.
Assistant Vice President for Research
Dean of the Graduate School & International Admissions

ABSTRACT

Rogers, Justin Leslie. *A Comparison Of Multivariate Methods For Measuring Change From Pretest To Posttest*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2011.

Three multivariate methods for measuring change from pretest to posttest are compared with respect to statistical power over various levels and combinations of effect size, alpha level, sample size, number of dependent variables, number of significantly different dependent variables, correlation between corresponding pretest and posttest scores, and correlation between unrelated pretest and posttest scores. The method utilizing posttests as the dependent variables and pretests as covariates was found to have superior statistical power in the majority of the scenarios examined. However, there were scenarios where the method utilizing change scores as dependent variables and the method utilizing only posttests as the dependent variables displayed greater power. Using results from the Monte Carlo simulations, comparisons are presented that reveal the conditions under which each of the three multivariate methods displayed greater statistical power than the other two. In addition to the immediate implications of the current study, suggested future avenues of research that could expand upon the current findings are discussed.

ACKNOWLEDGMENTS

The author wishes to thank the many people who played an invaluable role in the completion of this dissertation. First, I would like to thank Dr. Mundfrom, Dr. Perrett, Dr. Schaffer, and Dr. Heiny for their guidance throughout this process and for all that they have taught me during my time at University of Northern Colorado. Most importantly, I would like to thank my wife, Sarah, for her unconditional love, encouragement, and patience during this process. I would also like to express my deepest gratitude to my father and mother for all of the love and support they have given me throughout my life, as well as for all of the wisdom they have imparted to me over the years. Also, a great deal of thanks is owed to my many friends and family who always believed in me during these years. Last but not least, I would like to thank Hope Knuckles for encouraging me and keeping me motivated and Rose Grandy for her helpfulness and amazing SAS® skills.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

# CHAPTER I

## INTRODUCTION

A pretest measurement of the dependent variable is often captured in an independent two-group study design to more precisely evaluate the impact of a treatment upon a posttest measurement. In randomized experiments, the purpose for collecting both pretest and posttest scores for the same dependent variable often lies in the intuitive meaning of the subtraction of the pretest from the posttest. The researcher is interested in the amount of change for each subject; a subtraction of the pretest from the posttest reflects that amount of change for each subject, and therefore a change score per subject is a logical choice for the dependent variable (or outcome measure). Although the same motivation exists in quasi-experimental designs (non-randomized group membership), an additional and often problematic reason sometimes underlies the use of a change score. In this case, an adjustment is required because the two comparison groups (e.g., a treatment group and a control group) are not, on average, the same in value on the dependent variable prior to the treatment. Without an adjustment for this initial difference, conclusions could be misleading in that baseline group differences might simply carry over after the treatment. It has been argued that by subtracting the pretest from the posttest score that the two comparison groups have been equalized at baseline on the dependent variable (Lord, 1967).

Why the use of the change score to make this adjustment is problematic, especially in a quasi-experiment, is discussed later. The immediate point is that problems caused by change scores within a quasi-experimental setting have left the change score, regardless of the experimental design (even if randomized), with a somewhat tarnished reputation. Indeed, many influential articles, such as Lord (1967), Cronbach and Furby (1970), and Linn and Slinde (1977), have been written over the years criticizing the use of change scores. As Maxwell and Howard (1981) note, "An unfortunate by-product of these articles seems to have been the creation of the belief among many researchers that the use of change scores is universally misleading and therefore should be avoided at all costs" (p. 747).

In a broad sense, the current study adds to a growing number of others such as Maxwell and Howard (1981), Zimmerman and Williams (1982), and Allison (1990) that attempt to resurrect the change score. Indeed, it is quite useful and researchers need not avoid it if care is taken to address the problems so vigorously pointed out by Cronbach and Furby (1970) and Linn and Slinde (1977). In the narrow sense, this study is about using the change score in randomized experiments that require a multivariate array of dependent variables, and hence are analyzed using multivariate analysis of variance (MANOVA) or multivariate of covariance (MANCOVA) to capture the full treatment effect. In one fashion or another, the remainder of this dissertation addresses this point.

## Historical Criticism of the Change Score

Although the change score has been criticized on a number of grounds (Kessler, 1977; Linn & Slinde, 1977; O'Connor, 1972), a key paper by Cronbach and Furby (1970) stands out as a classic assault on change scores. Kessler (1977) succinctly summarizes the critical argument of the paper, pointing out that change scores based upon imperfectly measured components (i.e., the pretest and posttest measurements) are even less reliable than their individual components. A review of the Cronbach and Furby (1970) paper reveals many of the reasons that one might not want to use the change score as an outcome variable, a number of corrective actions that are possible if the change score is used, and what alternatives are available that preclude the need for the change score altogether. The despair over change scores that Cronbach and Furby exhibit is captured in their opening remarks:

> "Raw change" or "raw gain" scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement. Although the unsuitability of such scores has long been discussed, they are still employed, even by some otherwise sophisticated investigators. (p. 68)

An overview of Cronbach and Furby's (1970) position forms a backdrop and context within which the current study rests. First, the diminished reliability of raw change scores is addressed by showing how such scores can be modified so that future investigators who use them (regardless of advisability) will do so with less error. Relying on earlier presentations by Lord (1956, 1958, 1963) and McNemar (1958), Cronbach and Furby present several methods whereby the investigator can more accurately estimate the true change score, each better than the former, using regression

models. The final and superior method involves covariates (in addition to the pretest score) that are thought to correlate with the estimation of a true score. In all instances, the procedures provided by Cronbach and Furby require estimates of the reliabilities of the pretest and posttest measurements, the variances for these measurements, and the covariance between these measurements (Linn & Slinde, 1977). Separate formulas are presented for situations where uncorrelated (or independent) pretest and posttest scores are expected and situations where correlated errors of measurement (linked observations) would be suspected. Finally, these authors conclude by presenting additional alternative estimators that utilize residual scores around the pretest score to posttest score regression line.

Second and more importantly, Cronbach and Furby (1970) discourage the use of the corrected gain scores they present, arguing that alternative analysis strategies that do not rely on change scores should be used. They then match analysis methodologies that avoid the use of change scores to distinct research settings. Of great importance to Cronbach and Furby is that these methodologies do not use change scores and actually make the need for them unnecessary.

## Essential Limitations of the Change Score

Subsequent to Cronbach and Furby's (1970) classical presentation on the topic, Allison (1990) has made the point that the foundational problems with change scores are essentially twofold, and it is because of these two reasons that warnings about change scores have come about. The first reason involves the issue of reliability, which was the main motivation behind Cronbach and Furby's classic presentation.

The second is the closely related problem of regression toward the mean (also referred to as regression effects in the literature). This phenomenon arises from the idea that individuals who score high on the pretest tend to score lower (or move down as Allison referred to it) on the posttest, and individuals who score low on the pretest tend to score higher (or move up) on the posttest. Therefore, individuals with more extreme (very high or very low) pretest scores have a tendency to obtain less extreme posttest scores.

The first problem, in its most fundamental form, is noted by Kessler (1977) and summarized by Allison (1990): "Change scores tend to be much less reliable than the component variables" (p. 94). To illustrate this point, Allison notes that in the case where the pretest ($Y_1$) and posttest ($Y_2$) scores are equally reliable and have the same variance, the reliability of the change score ($Y_1 - Y_2$) is simply

$$\frac{\rho_y^2 - \rho_{12}}{1 - \rho_{12}}$$

where $\rho_{12}$ is the correlation between $Y_1$ and $Y_2$, and $\rho_y^2$ is their common reliability. Allison then points out in reference to $\rho_{12}$, "If this correlation is positive (as it almost always is), then the reliability of the change score must be less than $\rho_y^2$, often much less" (p. 95).

To clarify the second point regarding regression toward the mean, Allison (1990) states,

> Because of the almost universal phenomenon of regression toward the mean from the pretest to posttest measurements, $Y_1$ will usually be negatively correlated with $Y_1 - Y_2$. Thus, individuals with high pretest scores will tend to move down on the posttest, while individuals with low pretest scores will tend to move up. Consequently, if $X$ (or any other variable) is correlated with $Y_1$, it will tend to have a spuriously negative relationship with $Y_1 - Y_2$ (Markus, 1980). For these reasons, methodologists in the social sciences have repeatedly warned against the use of change scores. (p. 95)

## An Early Example of the Ambivalence Towards Change Scores

Without a doubt, the sobering warning of the dangers inherent in the use of change scores has impacted research that might have otherwise thoughtlessly used them. However, the relevance of change scores has never been completely dismissed in the literature. Responses to the criticism of change scores have varied widely. Some responses have claimed that change scores simply should not be used (O'Connor, 1972) and advocated the use of experimental designs that avoid them (Cronbach & Furby, 1970). However, some responses have also included analyses that correct for them (Williams & Zimmerman, 1996), clarification of circumstances under which they escape the problematic status assigned to them (Zimmerman & Williams, 1982), and the description of selected circumstances that demand them (Maxwell & Howard, 1981). These points are discussed further in the next chapter. In general, change scores are much better understood now than was once the case, and over time the literature has come to present a more balanced view of their use.

One article of early interest pertaining to the ambivalence surrounding the use of change scores is that by Lord (1960). It demonstrates the agony inherent within this issue. Lord (1960) starts by noting that a simple analysis of covariance (ANCOVA)

with two treatment groups, one covariate (the pretest), and one dependent variable (the posttest) can be conceptualized as a simple *t*-test carried out on the posttest scores regressed back to a common value (zero) on the covariate (pretest). This analysis has been frequently used to equate the treatment and control groups on the pretest score in quasi-experiments where baseline differences on the score exist. Lord (1960) then notes that if measurement error is associated with the covariate (pretest), even when the pretest and posttest are perfectly correlated and should regress back to a common score when the covariate value is zero, the scatter of *x* values away from the regression line (due to error of measurement) will force the treatment and control group to regress back to different values of the posttest score (even though the assumed perfect correlation should result in regressed scores to a common value). This observation led Lord (1960) to the conclusion that "the usual covariance analysis, which ignores the fallibility of *X*, will reach the erroneous conclusion that the difference between groups A and B on variable *Y* cannot be accounted for by the difference on variable *X*" (p. 309). That is, Lord (1960) concluded that the ANCOVA, when the covariate is measured imperfectly (contains error), can and often will lead to an unreliable conclusion. Indeed, Lord (1960) shows that a true difference between the treatment and control group can be obscured, as can a true equivalence between these groups. In an attempt to solve this issue, Lord (1960) presents a large sample covariance analysis approach that uses two pretest scores (rather than the typical single pretest score) to estimate and correct for the fallibility associated with the pretest measurement.

Interestingly, Cronbach and Furby (1970) state that when comparing treatment groups not formed at random, if ANCOVA is carried out, the comparison should be done using Lord's (1960) procedure. However, Cronbach and Furby also refer to this procedure as being "no more than a palliative" (p. 78). They go on to reinforce this point by quoting another paper written by Lord (1967) seven years after he first proposed the ANCOVA procedure that utilizes two pretest measures where he says, "there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups" (p. 305). This example characterizes the tone of many articles that oppose the change score as a valid assessment option.

**Experimental Design and the Analysis of Change Scores**

A major recommendation that develops from the scrutiny of change scores is that randomized experimental designs should be used if possible, and quasi-experimental designs should be avoided. Campbell and Stanley (1963) made this point early on in the debate. To this end, the literature divides in general along the lines of quasi-experimental methods and randomized experiments, with the positive role of change scores more pronounced in randomized studies. However, a great deal of work also exists that clarifies the conditions under which change scores can play a beneficial role in quasi-experimental designs. Interestingly, relatively little information exists concerning the use of change scores when multivariate statistics are required, particularly if the underlying design is quasi-experimental. It became more evident that a

primary objective of the current effort was the extension of knowledge concerning the use of change scores in multivariate randomized designs.

In general, the analysis of change scores can be considered within the context of the following four designs: univariate randomized experiment, univariate quasi-experiment, multivariate randomized experiment, and multivariate quasi-experiment. Most of the discussion lies within the realm of the univariate case. Little insight is available for multivariate analyses, and discussion appears to be non-existent for quasi-experimental designs, with only a limited discussion pertaining to randomized multivariate designs. However, it becomes evident in what follows that change scores can play an important and legitimate role in multivariate randomized designs.

**Univariate Randomized Design**

Confusion has existed concerning which of four different methods are best in experiments with both pretest and posttest scores available. The four common approaches are the analysis of posttest only, the analysis of the posttest with the pretest as a covariate, the use of change scores as the dependent variable, and the inclusion of the pretest and posttest score as a repeated measures factor in a two-way analysis of variance (ANOVA) with one repeated measure. Unfortunately, as Maxwell and Howard (1981) point out, much of this confusion exists because applied researchers who are unfamiliar with the nuances of statistical models have failed to understand the impact of randomization upon the expected values of the underlying models for the three latter approaches that were just described. It is evident from the literature that the four approaches are mathematically similar within randomized experiments in that

all yield an unbiased test of treatment effect, although a consideration of the degrees of freedom available under each model for a statistical test of treatment may lead the researcher to prefer one method over another. In experimental designs where randomization has been used to create the treatment and control groups, the literature has shifted away from recommending that change scores not be used to a focus on whether their use weakens or enhances the statistical power.

**Univariate Quasi-Experimental Design**

The literature in this area discusses the liabilities inherent in the analysis of quasi-experiments using change scores. A great deal of attention is paid to the fact that the limitations of change scores rest upon the reality that the pretest and posttest measurements are nearly always imperfect, meaning that measurement error is present. In univariate quasi-experiments, this fact leads to issues caused by a lack of reliability in the change score and by the regression of change scores toward the mean. Both the problem of reliability and regression toward the mean can lead to false conclusions. The literature strives to make known specific instances when change scores are appropriate or desirable in quasi-experimental designs. One prominent feature of the literature concerns the use of ANCOVA in quasi-experimental settings. Although the ANCOVA often does not avoid the problems inherent in the use of change scores, corrections can be applied that improve the interpretability of covariance analysis used in quasi-experiments under certain conditions. The most recent articles take the position that both change score analysis and ANCOVA, depending on the prevailing circumstances of a given experiment, can be useful approaches to the analysis of

quasi-experiments. A number of articles, such as Allison (1990), Maxwell and Howard (1981), and Fitzmaurice, Laird, and Ware (2004) describe the conditions under which a given method might be preferred. More recently, authors such as Cribbie and Jamieson (2004) conclude that Structural Equation Models are the best solution for measuring change in quasi-experiments.

### Multivariate Randomized Design

As in the univariate case, both ANCOVA and analysis using change scores yield unbiased conclusions. However, Maxwell and Howard (1981) conjecture that unlike the univariate case, the use of change scores in multivariate true experiments could increase statistical power relative to MANCOVA and certainly could do so relative to the analysis of a posttest vector alone. As explained later, this observation forms the basis of the research presented in this dissertation.

### Multivariate Quasi-Experimental Design

As previously noted, the literature review did not find any discussion of the use of multivariate quasi-experiments involving change scores. Although it is not the focus of this paper, attention to the advantages and disadvantages of change scores and covariance analysis in quasi-experimental multivariate settings provides an important focus for future statistical research.

## Justification for This Study

A repeated issue throughout the literature on change scores concerns which of the following approaches to the analysis of a two-group randomized study is best. The three possible options that give unbiased tests of the treatment effect are ANOVA

applied to only the posttest score, ANCOVA employing the pretest as a covariate and the posttest as the dependent variable, and ANOVA with the change score as the dependent variable. The important issue is that of power. Maxwell and Howard (1981) as well as Delaney and Maxwell (1980) make this point very clear. Bock (1975) and Huck and McLean (1975) address the issue of power in the univariate case. Both papers found, as Maxwell and Howard (1981) summarize, that in general the ANCOVA "is the most powerful of the three approaches" (p. 749).

However, it is important to emphasize that the literature above was referring strictly to the univariate case where there is a single dependent variable, and not a vector containing multiple dependent variables. The MANOVA using only a posttest vector of dependent variables, the MANOVA using the pretest vector as covariates and the posttest vector as outcomes, and the MANOVA using a vector of change scores as outcomes, all provide unbiased tests of the main effect. This point was not lost to Maxwell and Howard (1981), when they raised the issue of power available to two of the three multivariate options just mentioned, namely the MANCOVA and the multivariate analysis applied to change scores. Although Maxwell and Howard do not consider the case of a multivariate analysis applied to posttest scores only, it is easy to see how this analysis also might have played a role in their thought process. They note that the analysis of change scores may be useful when,

> the design is a multivariate pretest–posttest design. For example, pretest scores on *p* measures may be obtained for each subject prior to an experimental manipulation. After the manipulation, scores are obtained for the same set of *p* measures. If subjects have been randomly assigned to groups, either multivariate analysis of covariance (MANCOVA) or a MANOVA on the *p* change scores tests the same null hypothesis of no treatment effect. The

> primary factor influencing which technique should be used is statistical power, which is a complex function of mean differences, sample sizes, number of variables, and covariance matrices. (p. 751)

This comparison brings up some interesting points. Maxwell and Howard (1981) go on to say that error sum of squares will typically be smaller with the MANCOVA model than with the MANOVA, however the degrees of freedom for the error term in the MANOVA model must be larger than the degrees of freedom for the MANCOVA model. This difference could mean that the MANOVA model will be more powerful in situations where the number of posttest scores is large relative to the sample size. Another point that they make is the fact that the MANCOVA model, by design, necessitates that each posttest measure must be adjusted for by each pretest measure. At times, this can make the results difficult to interpret. On the other hand, the MANOVA design using change scores adjusts each posttest using only the corresponding pretest.

The discussion by Maxwell and Howard (1981) points to an important issue and an important possibility, specifically that a MANOVA design applied to change scores may at times provide greater statistical power than a MANCOVA design applied to the same multivariate data. It is not a large step to also consider how both of these tests compare to a MANOVA applied to posttest scores only in terms of statistical power. As was noted previously, many different research fields attempt to measure change in some fashion. This study gives an important recommendation of how that analysis should be done.

## Purpose and Research Question

This study extends the work of Maxwell and Howard (1981), Bonate (2000), and Tu, Blance, Clerehugh, and Gilthorpe (2005) by comparing the following three statistical techniques: MANOVA applied to posttest scores only, MANCOVA utilizing the posttest vector as outcomes and the pretest vector as covariates, and MANOVA with change scores as the vector of outcomes. The following research question is addressed:

> Q When pretest and posttest scores are collected, how does statistical power under different sample sizes, effect sizes, numbers of dependent variables, and degrees of correlation within and between the pretest and posttest scores compare between a MANOVA that uses change scores (posttest minus pretest) as dependent variables, a MANOVA that uses only posttest scores as dependent variables and a MANCOVA that uses posttest scores as dependent variables and pretest scores as covariates?

## Limitations

This dissertation considered only the situation where the assumptions of multivariate normality, homogeneity of variance–covariance matrices, and linearity among all pairs of predictors exist. Therefore, the results of this study cannot be extrapolated to experiments when these assumptions are not met. This study only examined scenarios where the pretests are assumed to be equal between groups, thus the results found herein are not appropriate for studies where pretests are not assumed to be equal, such as a quasi-experimental design where group assignment was based on pretest scores. Also, only the two-group case was considered, so these results do not apply to studies that use three or more groups.

## Conclusion

This chapter has shown the need for further research in the area of multivariate pretest–posttest designs. Maxwell and Howard (1981) introduced the idea that using change scores as outcome variables may be preferable in multivariate designs relative to the use of posttest scores as outcome variables and pretest scores as covariates. Bonate (2000) performed a Monte Carlo simulation examining the performance of 11 different methods for measuring change, including change scores and ANCOVA with posttest scores as the dependent variable and pretest scores as the covariate, but all of the comparisons made were within the univariate realm. Tu et al. (2005) performed a Monte Carlo simulation as well that examined posttest only, change scores, percent change, ANCOVA with the posttest as the dependent variable and pretest as the covariate, a random effects model, and MANOVA. However, the MANOVA model differed substantially from the three models under examination in this dissertation. They used the pretest score and the corresponding posttest score as the two dependent variables. This model is not intuitive and is not a natural extension of the models ordinarily used in the univariate case. In short, a formal comparison of the three most common univariate models applied to change scores, when generalized to the multivariate case, does not yet exist. As it stands, a multivariate analysis has not been done up to this point examining the multivariate situation previously described. This dissertation examined, under varying conditions, which type of pretest–posttest multivariate analysis is preferable with respect to statistical power. The specific conditions under which these comparisons occur are discussed in Chapter III.

## Terminology

The following terminology will be used in this dissertation:

**Change score.** The difference obtained from subtracting the pretest score from the posttest score is a change score (also referred to as a difference score, gain score, or growth score in the literature).

**Effect size.** This is the difference between the means of two groups for a given variable expressed in terms of standard deviation units.

**Power.** This is the probability that a statistical test will correctly reject the null hypothesis when a statistically significant difference between two groups exists.

**Pretest–posttest experimental design.** This is an experiment comparing two groups using paired data where a subject or experimental unit is measured at either two separate points in time or at the same time under two different testing conditions. The first measurement is referred to as the pretest or baseline, and the second as the posttest. The researcher is interested in determining whether or not a statistically significant difference exists between the pretest and the posttest or if two or more groups have significantly different measurements between pretest and posttest.

# CHAPTER II

# REVIEW OF LITERATURE

As was previously discussed, this study extends the work of Maxwell and Howard (1981), Bonate (2000), and Tu et al. (2005) to the multivariate realm in order to compare and contrast three statistical methods for examining pretest–posttest designs.

This chapter is broken up into the following three sections:

1. The history and debate of how to measure change in pretest–posttest study designs.
2. A brief history and background of MANOVA and MANCOVA.
3. A closer look at the Maxwell and Howard (1981) paper and the Monte Carlo simulation studies performed by Bonate (2000) and Tu et al. (2005) for univariate pretest–posttest designs.

## History and Debate of How to Measure Change Over Time in Pretest–Posttest Designs

In Chapter I, the debate over best practices for the analysis of change over two time points—pretest and posttest—was introduced. It was shown that debate over the advantages and disadvantages of using change scores often has been fueled by research where change scores were used to allegedly overcome baseline discrepancies between

a treatment and control group in a quasi-experimental setting where random assignment had not been used. A consideration of the general debate exposed various viewpoints and themes. These embraced a scattering of reasons not to use change scores (Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1967; O'Connor, 1972), reasons to use change scores (Allison, 1990; Maxwell & Howard, 1981; Zimmerman & Williams, 1982; Zumbo, 1999), corrective actions that may improve change scores (Cronbach & Furby, 1970; Lord, 1960; Williams & Zimmerman, 1996), and analyses or experimental designs that avoid change scores altogether (Campbell & Stanley, 1963; Cribbie & Jamieson, 2004; Cronbach & Furby, 1970).

In general, considerations surrounding the use of change scores were threefold. The first concern focused on the decreased reliability of a change score relative to each of the two scores comprising it. The second addressed the closely related phenomenon of regression toward the mean over time when the pretest is measured with imperfect reliability. The third concern, forming the emphasis of this dissertation, was that of available statistical power in true experiments that employ randomization.

The following articles by Gottman and Krokoff (1989, 1990) and Woody and Costanzo (1990) illustrate the sometimes heated discussions that have occurred over how one should measure change. Although the quasi-experimental study used in this illustrative study differs from that of a true experiment, which is assumed in this dissertation, these articles are representative of the confusion and debate surrounding the use of change scores.

Gottman and Krokoff (1989) performed a study with the goal of predicting marital satisfaction from micro-component measurements of anger, contempt, fear, sadness, and whining as measured by the Marital Interaction Coding System, the Couples Interaction Scoring System, and the Specific Affect Coding System. Marital satisfaction was determined by the use of the Locke–Wallace (Locke & Wallace, 1959) and the Locke–Williamson (Burgess, Locke, & Thomes, 1971) scales and was measured at baseline and three years later. Regression analysis was used to assess the predictive value of the micro-components. Specifically, the micro-components of the husband and wife were regressed on a change score consisting of the marital satisfaction posttest score minus the marital satisfaction pretest score. A major conclusion was as follows:

> Wives who are positive and compliant fare better in terms of their husband's concurrent negative affect at home and concurrent marital satisfaction, but the marital satisfaction of these couples deteriorates over time. On the other hand, the stubbornness and withdrawal of husbands may be most harmful to the longitudinal course of marital satisfaction. In terms of specific emotions, the marital satisfaction of wives improves over time if wives express anger and contempt during conflict discussions but declines if the wives express sadness or fear. For husbands, only whining predicts change in marital satisfaction over time, and it predicts the deterioration of both partners' marital satisfaction. Thus, we cannot say that the same negative affects are equally positive or negative, in a longitudinal sense, for husbands and wives. In terms of recommendations for marriage, our results suggest that wives should confront disagreement and should not be overly compliant, fearful, and sad but should express anger and contempt. Husbands should also engage in conflict but should not be stubborn or withdrawn. Neither spouse should be defensive. (Gottman & Krokoff, 1989, p. 51)

Of interest here is that in this quasi-experimental study, a change score served as the dependent variable and that a heated debate over its use, as shown below, soon arose.

Shortly after the publication of that article, Woody and Costanzo (1990) published a strongly worded objection to Gottman and Krokoff's (1989) use of change scores in the marital satisfaction study. Their response consisted of arguments founded on "traditional psychometric concerns" (p. 499) that were assured by measurement error (less than perfect reliability) in the pretest, posttest, and micro-component scores, as well as the problem of regression towards the mean. Their first point, focusing on traditional psychometric concerns, is captured in the following statement:

> Gottman and Krokoff measure 3-year change in marital satisfaction by subtracting each initial score from the score obtained 3 years later. They then correlate the interaction variables with these difference scores. Now, the correlation of a variable *v* with a difference score (*a*–*b*) may be expressed as (Cohen & Cohen, 1983, p. 416).
>
> $$r_{v(a-b)} = \frac{r_{va}SD_a - r_{vb}SD_b}{\sqrt{SD_a^2 + SD_b^2 - 2r_{ab}SD_aSD_b}}$$
>
> To see the implications of this equation, let us assume that $SD_a = SD_b$. Then Equation 1 reduces to
>
> $$r_{v(a-b)} = \frac{r_{va} - r_{vb}}{\sqrt{2 - 2r_{ab}}}$$
>
> From this equation we can see that the correlation of a variable *v*, such as an interaction measure with Time 1 scores on marital satisfaction, *b*, can make a substantial inverse contribution to the correlation of *v* with the 3-year change score (*a*–*b*). (pp. 499-500)

Woody and Costanzo (1990) drew the conclusion that the negative correlations cited by Gottman and Krokoff (1989), which underlie important aspects of their reported findings, were due to a statistical artifact. It should be noted that even if *a* and *b* were simply two measurements of a constant attribute and differed from one

another only through lack of perfect measurement reliability, that Woody and Costanzo's observation would hold. Thus, their argument is driven in part by the lack of reliability that will invariably exist between a pretest and a posttest measurement.

The second concern that Woody and Costanzo (1990) address is regression to the mean across time. They state:

> It is highly likely that the scores of both highs and lows will regress toward the mean at Time 2 (because the correlation of Time 1 with Time 2 marital satisfaction is substantially less than one). This means that at Time 2 the scores of the lows will have increased, whereas those of highs will have decreased. As a result, the variance of marital satisfaction will be *less* for Time 2 than for Time 1. That is, referring back to Equation 1, the extreme-groups nature of the sample will make $SD_a$ less than $SD_b$. This reduction of variance from Time 1 to Time 2 worsens the confounding of the difference scores, ($a$–$b$), with the initial scores, $b$. To see this, note that the numerator of the expression in Equation 1 is
>
> $$r_{va}SD_a - r_{vb}SD_b.$$
>
> The contribution of each correlation, $r_{va}$ and $r_{vb}$ to $r_{v(a-b)}$ is *weighted* by the associated standard deviation of $a$ and $b$. Hence, $r_{vb}$ (the correlation of the interaction variable with Time 1 marital satisfaction) contributes more heavily to $r_{v(a-b)}$, *again* in an inverse fashion. (p. 500)

The main point of these criticisms is that the inverse correlations that ground the substantive conclusions made by Gottman and Krokoff (1989) are promoted by statistical artifacts stemming from lack of reliability (at least in part) and regression towards the mean. Woody and Costanzo (1990) go on to offer two solutions, one involving structural equation models and the other the ANCOVA. Regarding the latter, they suggest as a partial solution to the problem, the use of a residualized scores analysis, which is simply the use of the pretest score as a covariate in an ANCOVA containing the pretest and one or more of the micro-component measurements and

posttest score as the dependent variable. However, Woody and Costanzo admit that this approach "raises its own issues" (p. 500), but nevertheless believe that many methodologists would consider it to be "an important step in the right direction" (p. 500).

In what amounts to a third full journal article dedicated to this debate, Gottman and Krokoff (1990) offered a detailed rebuttal to the above criticisms. They review five methods that might be used to analyze the data—four of which embrace residualized scores (utilize an ANCOVA method) and one that embraces change scores. Gottman and Krokoff (1990) algebraically manipulate each formula of the five methods, and conclude that "the suggestion made by Woody and Costanzo (1990) on the issues of statistical approaches to the study of longitudinal change is no real improvement in the statistical sense" (p. 503).

Gottman and Krokoff (1990) also address the criticism of Woody and Costanzo (1990) that refers to regression toward the mean. They state that "regression toward the mean does not imply that the variance decreases from initial to final score" (p. 502). They also make the following important point:

> The problem of regression to the mean is exacerbated by a distribution more humped near the mean that at its tails. The problem of regression to the mean is reduced when the distribution is rectangular (i.e. when each part of the sampling distribution is equally likely). Because this is the case, oversampling the tails of a distribution (that is, the oversampling of extreme groups) forces the distribution to be more rectangular and reduces regression to the mean. This was the logic of our sampling procedure (Gottman and Krokoff, 1989), and our distribution is indeed nearly rectangular. Hence, rather than exacerbating the problems, as Woody and Costanzo contend, the oversampling of the tails is actually at the core of solving the problem of regression toward the mean. Thus it is not the case that extreme groups may "exacerbate this contamination" (p. 500), as Woody and Costanzo suggest. (p. 502)

As a final point, Gottman and Krokoff (1990) claim that the change score most accurately captures their intent. Lord (1967) states that some people,

> assert that deviation from the regression line is the real measure of change, and that the ordinary difference between initial and final measurement is not a measure of change. This can hardly be correct. If certain individuals gained 300 ounces, this is a definite fact, not a result of an improper definition of growth. (p. 23)

Gottman and Krokoff (1990) mirror this idea when they describe the use of change scores as being "clear and simple in the sense that it has a precise interpretable physical meaning. It is, quite simply, the amount of change. The deviation from a regression line is a more complex statistic to interpret" (p. 504). Finally, in contrast to Woody and Costanzo's (1990) sentiment that "the prediction of raw change may be devoid of interest" (p. 500), Gottman and Krokoff (1990) go on to point out that the prediction of raw change was exactly what their research was interested in.

By reviewing the Gottman and Krokoff (1989) article and the exchanges that followed (Woody & Costanzo, 1990, and Gottman & Krokoff, 1990), the importance of the three earlier noted considerations that underlie the use of change scores (measurement error, regression toward the mean, and the use of randomized versus quasi-experimental design) may be seen. Measurement error and regression toward the mean underlie both the criticism offered by Woody and Costanzo (1990) and the response to it by Gottman and Krokoff (1990). Although not directly raised by either group of authors, the failure to examine the research question using a randomized design allows the debate to exist. Rather than focusing on the statistical power available to different statistical approaches applied to unbiased estimators made

possible by randomization (as will be discussed later), a great deal of energy was exerted toward the creation of the proper adjustment to compensate for this fundamental limitation in research design. In what immediately follows, literature bearing on various implicit and explicit issues raised by this illustrative debate are examined in greater depth.

The research question (or hypothesis) is actually very important in deciding whether one should use ANOVA with change scores or ANCOVA with posttest scores as the dependent variable and pretest scores as the covariate. By claiming to have discovered a paradox, Lord (1967) seems to have caused much confusion and led many researchers astray by failing to understand what is being tested by each of the two methods he imagines might be used to evaluate a research question (Fitzmaurice et al., 2004). He incorrectly assumes that both methods are testing the same hypothesis, and therefore finds it paradoxical that situations exist in which the two methods could come to completely different conclusions. The hypothetical example Lord (1967) uses to illustrate this paradox is a measurement on males and females at two different time points when a diet program is started at their university. Both the group of males and group of females exhibited the exact same weight gain, even though the males weighed more than the females at the start of the study. Using two different approaches to analyze the data from his hypothesized experiment, Lord (1967) determined that change scores did not detect a significant difference between the groups, but ANCOVA did detect such a difference. Lord (1967) concludes that "confused interpretations may arise from such studies" (p. 305) and in his opinion,

"there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled pre-existing differences between groups" (p. 305). Cronbach and Furby (1970) echoed Lord's (1967) sentiment when they recommend that "investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways" (p. 80). This matter has come to be known as Lord's paradox.

The problem is that Lord (1967) failed to recognize that the two methods answer different research questions. Fitzmaurice et al. (2004) say that using a change score "addresses the question of whether the two groups differ in terms of their mean change over time" (p. 124). In contrast, ANCOVA, using posttest as the dependent variable and the pretest as a covariate, tests whether a difference exists between the posttest scores of two or more groups after adjusting for differences that may have existed at the pretest. That is, ANCOVA tests whether two or more groups improved (or declined) at the same rate starting from the same mathematically determined baseline mean value. Fitzmaurice et al. say that ANCOVA "addresses the question of whether an individual belonging to one group is expected to change more (or less) than an individual belonging to the other group, *given that they have the same baseline response*" (p. 124). This analysis contrasts with the absolute amount of change, regardless of baseline, that is the focus of a change score analysis. Thus, Fitzmaurice et al. conclude that the choice to use the change score method or the ANCOVA method should depend upon the research question.

Although it was not his most well known paper in regard to the measurement of change, Lord's (1956) article may have sparked the debate over change scores by introducing two of the points of contention previously noted: measurement reliability and regression toward the mean (also referred to as regression effect). In calculating the reliability of change scores, Lord (1956) makes the assumption that the variance of the posttest score will be equal to the pretest score. As shown in the first chapter, Allison (1990) gives an illustration of how the reliability of change scores must be low when the assumptions by Lord (1956) are followed. From his derived formulas, Lord (1956) concludes that "[d]ifferences between scores tend to be much more unreliable than the scores themselves" (p. 429).

McNemar (1958) pointed out that Lord (1956) is actually incorrect in assuming that the variances of the pretest and posttest scores will be equal. He felt that Lord's (1956) assumption was unrealistic and too restrictive for what is seen in typical research. The assumption of equal variances is actually untenable when considering most situations in which growth would be measured, such as mental or educational growth. He goes on to show that the reliability of gain scores is much better when an assumption of equal variances is not present. However, as evidenced by the appearance after his publication of articles (Cronbach & Furby, 1970; Linn & Slinde, 1977) that utilized Lord's (1956) argument to oppose change scores, McNemar's attempt to clarify the debate of change scores with respect to reliability was in large part unnoticed.

Cronbach and Furby (1970) felt that both Lord (1956) and McNemar (1958) had sidestepped the "philosophically troublesome question, Are pretest and posttest 'measuring the same variable'?" (p. 69). Linn and Slinde (1977) explain that the best way to obtain better reliability of change scores is to have low correlation between the pretest and posttest scores. If this is the case, though, they question whether the "pre- and postmeasures are getting at the same construct, which would seem to be a prerequisite for the difference score to be interpreted as an index of growth" (pp. 123–124). Therefore, they felt it was risky to make important decisions based on change scores because they presume researchers will either encounter low pretest–posttest correlation or apparently low measurement reliability. Linn and Slinde went on to show that the reliability for the ANCOVA method is better than that of the change score method, but is still disappointingly low when the correlation between pretest and posttest scores is high.

Overall and Woodward (1975), Zimmerman and Williams (1982), and Rogosa (1988) defended change scores with regard to reliability. Overall and Woodward demonstrated that the statistical power of change scores is actually maximized when the subsequent reliability is zero, and therefore not a valid argument against them. However, Zimmerman and Williams and Rogosa make an even more convincing argument, saying that the assumptions used by Lord (1956) and Linn and Slinde (1977) are incorrect. Zimmerman and Williams pointed out that it is not necessarily the case that the reliabilities of the pretest and posttest are always equal. They also pointed out that McNemar (1958) was correct in saying that the variances (and

therefore the standard deviations) will most likely be unequal. Rogosa noted that the variance of the posttest will oftentimes be greater than that of the pretest. Zimmerman and Williams demonstrate that the reliability of the change scores are consistently high when their assumptions of unequal reliabilities and unequal variances are true. Rogosa extended the argument, saying that previous authors such as Linn and Slinde (1977) had confused the observed correlation with measurement error and the true correlation (which is free of error) with the assumption that the variance of a measure remains stable over time. Rogosa deduced that this confusion has lead to incorrect conclusions and has misled researchers when, in fact, "the difference score is an unbiased estimator of true change" (p. 180).

The other issue that Lord (1956) introduced was regression toward the mean. Cronbach and Furby (1970), O'Connor (1972), and Linn and Slinde (1977) attacked the use of change scores using regression toward the mean as the basis of their argument. They argued that this effect occurs due to the negative correlation between the pretest score and the change score. O'Connor explains that the "correlation between change and initial status is biased in a negative direction by errors in the pretest because the pretest error is also present in the change score but with the opposite sign" (p. 74). Therefore, these authors believed that the results from an ANOVA with change scores would be biased due to this regression effect. It is of interest to note that this phenomenon is typically attributed to situations where randomization has not been used to create group membership (Maxwell & Delaney, 2004). In randomized controlled study designs, both the change score method and the

ANCOVA method are unbiased because of the assumption that no baseline differences exist between the groups (Oakes & Feldman, 2001).

Zimmerman and Williams (1982) pointed out that the correlation between the pretest score and the change score can actually be positive or zero and not just negative. Again, this incorrect premise by Cronbach and Furby (1970) and Linn and Slinde (1977) stems from the incorrect assumption that the variances of the pretest and posttest scores are equal (Rogosa, 1988). Regression toward the mean only occurs when the variances of pretest and posttest scores are equal. Rogosa considered the occurrence of equal variances to be a very rare event as variance typically increases over time. He also pointed out that even when stable variances do occur, using the ANCOVA method does not necessarily avoid the problem. Finally, Maris (1998) states, "regression toward the mean is not a reason for not using the gain score estimator" (p. 325). Maris regarded regression toward the mean and a biased change score estimator as,

> two aspects of the same data pattern, and there is no logical relation between the two phenomena. In particular (a) regression toward the mean does not imply that $\hat{\tau}^{gain}$ is biased, and (b) the absence of regression toward the mean does not imply unbiasedness of $\hat{\tau}^{gain}$. (pp. 322–323)

Here, note that Maris used $\hat{\tau}^{gain}$ to represent the change score estimator.

Many authors (Linn & Slinde, 1977; Lord, 1956; O'Connor, 1972) have detailed examples of situations that could arise and lead to bias from using change scores. Ironically, both Allison (1990) and Oakes and Feldman (2001) pointed out that in such non-randomized situations, the change score is actually less biased than the ANCOVA method, if it is biased at all. On the other hand, Fitzmaurice et al. (2004)

showed that when differences exist in the pretest scores between groups, the ANCOVA method can lead to biased or misinterpreted results. In such a situation, covariates can introduce spurious relationships between the variable denoting group membership and the posttest. The researcher could come to the conclusion that there is no difference between groups when one truly existed, simply because the covariate explained away the meaningful group differences. Furthermore, Fitzmaurice (2001) shows that in situations with nonequivalent groups, the ANCOVA method often does not answer the intended research question.

In the discussion concerning change scores immediately above, it was noted by way of reference to Maxwell and Delaney (2004) and Oakes and Feldman (2001) that neither ANOVA using change scores nor ANCOVA using posttest scores as the dependent variable and pretest scores as the covariate provide biased estimates in randomized trials. This concept is of great importance to the basis of this dissertation, which assumes the setting of a randomized controlled trial. Without the concern of a biased estimation, statistical power becomes the focus. Oakes and Feldman explain that studies lacking sufficient statistical power can lead to incorrect conclusions and waste resources, doing more harm than good. Therefore, it is important to use the test statistic that provides the greatest amount of statistical power.

Many studies (Bonate, 2000; Fitzmaurice et al., 2004; Maxwell & Delaney, 2004; Maxwell & Howard, 1981; Tu et al., 2004) have compared the statistical power between the ANCOVA method and the change score method in univariate randomized controlled trials. In contrast to the disagreement around the use of change scores in the

quasi-experimental setting (summarized above), these authors largely (but not completely as described later) arrived at a common conclusion.

It is generally thought that ANCOVA has an advantage in terms of power when compared to change scores. Bonate (2000) and Tu et al. (2004) used Monte Carlo simulation studies to examine statistical power, and both sets of authors concluded that the ANCOVA method exhibited an advantage with regard to existing statistical power. Maxwell and Delaney (2004) conjectured that diminished power in the case of change scores is due to the fact that the error variance of the ANCOVA method tends to be smaller than the error of the change score method. They argued that this fact gives the change score method less power and less precision than ANCOVA because, potentially, a smaller amount of error around the regression line exists in ANCOVA. Thus, change scores may miss a difference that ANCOVA is able to detect due to the improvement in power and precision.

There are also two plausible exceptions to the power comparisons that generally have been agreed upon in the literature. Oakes and Feldman (2001) believed that "the common assumption that ANCOVA models are more powerful rests on the untenable assumption that pretests are measured without error. In the presence of measurement error, change–score models may be equally or even more powerful" (p. 18). Also, Maxwell and Delaney (2004) suggested that with small samples in the two-group case, change scores could potentially offer more power because there is one less parameter to estimate than in ANCOVA.

In the case of univariate analysis conducted in a randomized controlled study, there is, for the most part, a consensus concerning the available power when ANOVA with change scores is compared to ANCOVA using posttest as the dependent variable and pretest as the covariate. The ANCOVA method seems to consistently display greater power in simulation studies when compared to the change score method. However, the available power still remains to be decided in the multivariate realm, since there has not yet been a study that has taken on this task. The goal of this dissertation was to extend previous work just described into a multivariate realm.

## History and Background of MANOVA and MANCOVA

Hershberger (2005) states that MANOVA was built on the foundations of Karl Pearson's chi-square distribution that was derived in 1900, W.S. Gossett's (student's) *t* distribution that was derived in 1902, and R.A. Fisher's ANOVA that was introduced in 1923. Hershberger notes that Fisher's ANOVA was derived to test population differences on $p = 1$ dependent variable, but the interest was soon turned to testing population differences on $p > 1$ dependent variables. Within a decade, Wilks (1932) extended Fisher's (1922) application of maximum likelihood estimation from the comparison of multiple groups on one dependent variable to multiple groups on multiple dependent variables simultaneously based on the generalized likelihood-ratio (LR). However, it was not until 1946 that the actual term, multivariate analysis of variance (MANOVA), was coined by Roy (1946).

A detailed description of the derivation of MANOVA can be found in numerous textbooks, such as Johnson and Wichern (2002). Briefly, Wilks (1932) assumed a

multivariate normal probability density function and a likelihood of a sample from this distribution as $L_0$ for the null hypothesis ($H_0$) and $L_1$ for the alternative hypothesis ($H_1$). The ratio of $L_0/L_1$ can be used to test the null hypothesis that all *k* samples are drawn from the same population versus the alternative hypothesis that at least one of the *k* samples is drawn from a different population. Tabachnick and Fidell (2001) explain that the test statistic derived is similar to ANOVA, since ANOVA uses a ratio of variances, or mean squares, to test main effects and interactions. The numerator represents the between-groups variance, and the denominator represents the total variance. In MANOVA, the determinants of the cross-products matrices are analogous to the mean squares, and the ratios of determinants test the main effects and interactions. Wilks's lambda thus follows the general form of

$$\Lambda = \frac{|S_{error}|}{|S_{effect} + S_{error}|}$$

In this formula, $|S_{error}|$ is the determinant of the error cross-products matrix. and $|S_{effect} + S_{error}|$ is the determinant of the sum of the error and effects cross-products matrices.

In order to evaluate Λ, Bartlett (1939) proposed an approximation based on the $\chi^2$ distribution. Other approximations were later derived. Rao (1952) developed an *F* statistic that better approximated the Λ cumulative probability densities than the chi-square distribution. Since Rao's book, other commonly used test statistics based on the *F* distribution were developed such as Hotelling's trace, Pillai's trace, and Roy's greatest common root criterion (Tabachnick & Fidell, 2001). However, according to

Haase and Ellis (1987), all of these test statistics are identical to Wilks's lambda in situations where there are only two groups being compared.

Finally, Hershberger (2005) describes and illustrates that,

> as ANOVA can be extended to the analysis of covariance (ANCOVA), MANOVA can be extended to testing the equality of group means after their dependence on other variables has been removed by regression. In the multivariate analysis of covariance (MANCOVA), we eliminate the effects of one or more confounding variables (covariates) by regressing the set of dependent variables on them; group differences are then evaluated on the set of residualized means. (p. 867)

Bartlett (1947) is credited as the first person to publish an analysis utilizing MANCOVA.

## A Closer Look at Key Sources

In the final section of this chapter, key sources that lay the groundwork for the present study were highlighted and more carefully considered. Maxwell and Howard (1981), in their defense of the use of change scores, were the first to suggest that change scores might be useful in multivariate analyses. Maxwell and Howard state that in univariate randomized pretest–posttest study designs, ANCOVA using the posttest as the dependent variable and the pretest as a covariate is a more powerful test than an ANOVA using the change score as the dependent variable. However, they point out that change scores are still valid in randomized controlled trials because using them is mathematically equivalent to a repeated measures analysis and still provides unbiased results. Maxwell and Howard go on to say that there are at least two other situations where change scores might be the preferred method of analysis for

pretest–posttest studies: multivariate analyses (which is the focus of this dissertation) and when response–shift bias is present.

Maxwell and Howard (1981) describe the logic behind their assertion that change scores could be superior in multivariate pretest–posttest settings. They explained that if subjects are randomly assigned to groups, a MANOVA with $p$ change scores or a MANCOVA with $p$ posttest scores and the $p$ corresponding pretest scores as covariates are both appropriate methods for testing the same null hypothesis of no treatment effect. They state that the primary determining factor for choosing which one of these two methods to use should be statistical power, "which is a complex function of .mean differences, sample sizes, number of variables, and covariance matrices" (p. 751).

Maxwell and Howard (1981) compared the two multivariate methods based on the error sum of squares and the error degrees of freedom. They note that the error sum of squares for the MANCOVA model with posttest scores as the dependent variables and corresponding pretest scores as covariates were typically smaller than the MANOVA with change scores, just as in the univariate case. However, they point out that the error degrees of freedom must always be smaller for a MANOVA with change scores. Maxwell and Howard illustrate this idea by giving an example in the two group scenario: the error degrees of freedom is $n_1 + n_2 - p - 1$ in the MANOVA case, but $n_1 + n_2 - 2p - 1$ in the MANCOVA case. Finally, they conjecture that the smaller error degrees of freedom for the MANOVA with change scores could counteract the typically smaller error sum of squares for the MANCOVA with posttests as the

dependent variables and pretests as the covariates, allowing the MANOVA with change scores to have more power. However, they do not discuss any further under what conditions this concept may or may not be true.

In a Monte Carlo simulation study, Bonate (2000) examined 11 different methods for analyzing univariate two-group randomized controlled pretest–posttest studies. Of interest was which of the 11 methods were more powerful than others given different correlations between the pretest and posttest. Among the 11 methods studied were ANOVA with posttest only, ANOVA with change scores as the dependent variable, and ANCOVA with posttest as the dependent variable and pretest as the covariate, which are directly relevant to this dissertation.

To perform the Monte Carlo simulation, Bonate (2000) used $n = 10$ subjects in each of the two groups and $\alpha = 0.05$ to determine a statistically significant group difference. The correlation between pretest and posttest was systematically varied for effect sizes of 0, 1.0, 1.5, and 2.0. The correlation values used were 0, 0.25, 0.50, 0.75, 0.90, and 0.95. One thousand simulations were run for each combination of correlation and effect size.

Bonate (2000) made several observations of interest with regard to his simulation results. The percent of simulations to correctly reject the null hypothesis of no treatment effect increased as effect size increased, as expected. However, there were differences seen between some of the methods of analysis used. Bonate observed that, in general, when the correlation between pretest and posttest was less than 0.50, ANCOVA models had greater power than ANOVA models. Conversely, when the

correlation increased to 0.75 or above, ANOVA models with change scores as the dependent variable displayed slightly better power than ANCOVA with posttest as the dependent variable and pretest as the covariate. The power of the ANCOVA models tended to decrease as correlation increased, while the power of the ANOVA models remained relatively constant. On the other hand, the power of ANOVA with posttest only "dropped like a rock falling off a cliff" (p. 141) as the correlation between pretest and posttest increased. Bonate concludes that, in general, ANCOVA with posttest as the dependent variable and pretest as the covariate is the most powerful test.

Other recent literature has also touched on the idea of the use of multivariate analysis to measure change. Tu et al. (2005) performed a Monte Carlo simulation study examining six different methods for analyzing change in two-group randomized controlled pretest–posttest studies. Although their paper resided solely in the realm of univariate trials, it is important to note that one of the six analysis methods studied was a MANOVA.

Tu et al. (2005) compared a *t* test on posttest scores, a *t* test on change scores, a *t* test on percent change scores, an ANCOVA models with posttest as the dependent variable and pretest as the covariate, a random effects model (REM), and a MANOVA model with the pretest and posttest as the dependent vector. It is important to note that the *t* tests performed in their study are mathematically equivalent to a two-group ANOVA performed on the same variables (Rosner, 2000). Because their study either used univariate models or used MANOVA with the dependent vector comprised of the pretest and posttest score, the Tu et al. study is different than the current study. This

dissertation does not treat pretest scores as dependent variables in the MANOVA and MANCOVA models being examined; instead, the pretest scores are only utilized as covariates or in calculating the change scores. Therefore, the Tu et al. study differs from this dissertation because their paper only examines the univariate case (in the sense that their comparative conditions never include more than a single posttest variable), because unlike the current dissertation the pretest appears as an element in the dependent vector in the single multivariate model found among their comparison conditions, and because the single multivariate model in their study is always compared to a univariate model (never another multivariate model as in this dissertation).

Although these differences establish that the purpose of the study by Tu et al. (2005) is quite different than that of the current dissertation, their findings are not devoid of interest in the present setting. Indeed, they found a selected utility in the use of MANOVA rather than some of the five univariate models against which it was compared. Tu et al. found that their MANOVA method had greater power than change score, percent change score, and REM when the correlation between pretest and posttest ($\rho_{within}$) was low. However, when ($\rho_{within}$) was high, MANOVA was not as powerful compared to the other methods. This finding establishes the fact that MANOVA can, under certain conditions, provide greater power than a univariate test, even though the univariate and multivariate analyses used identical pretest and posttest scores (with the exception that one univariate model used only the posttest score). The finding that their multivariate model sometimes provided greater power than a univariate approach gives some justification for the comparison of different

multivariate models that involve both a pretest and posttest score, which is the focus of this dissertation.

The assertion that change scores may be the preferred and more powerful method compared to using posttests as the dependent variables and pretests as the covariates in multivariate pretest–posttest situations by Maxwell and Howard (1981), as well as the Monte Carlo simulation studies by Bonate (2000) and Tu et al. (2005), have laid the foundation for the current study. As will be seen Chapter III, many of the same independent variables that were systematically varied in the Bonate and Tu et al. Monte Carlo simulation studies are considered. The goal of this dissertation was to test Maxwell and Howard's assertion under many conditions by extending these previously studied univariate Monte Carlo simulations to the multivariate realm.

# CHAPTER III

# METHODOLOGY

The following research question will be addressed in this dissertation:

Q When pretest and posttest scores are collected, how does statistical power under different sample sizes, effect sizes, numbers of dependent variables, and degrees of correlation within and between the pretest and posttest scores compare between a MANOVA that uses change scores (posttest minus pretest) as dependent variables, a MANOVA that uses only posttest scores as dependent variables and a MANCOVA that uses posttest scores as dependent variables and pretest scores as covariates?

This dissertation helps to answer an open issue concerning the statistical power for these three models that was raised by Maxwell and Howard (1981). The issue of power has been systematically addressed in this dissertation using Monte Carlo simulations. Results of the Monte Carlo simulations were obtained after manipulating certain variables that could impact the outcome of the simulation. These variables were cross-classified so that each manipulated variable in the cross-classification scheme could be considered in light of the others. Note that these manipulated variables are called independent variables throughout this dissertation. In the present context, the term independent variable does not refer to the parameter in the statistical model that differentiates the control from the treatment group or any covariate in any one of the three multivariate models under study, but rather refers to the manipulated

dimensions described in this chapter that form the backdrop within which comparisons of available power are made.

Assuming a constant Type I error rate, the available statistical power (1 – *P*(Type II error)) was compared between MANOVA applied to change scores, MANOVA applied to posttest scores only, and MANCOVA using the posttest scores and pretest scores as dependent variables and covariates, respectively. A Monte Carlo simulation procedure, described in detail later, was used to calculate the available power for each multivariate model when differences are intentionally created between two multivariate normal distributions on the mean vectors of each. In the Monte Carlo simulation, each of the multivariate models listed above, when parameterized to capture the difference in mean vectors between the two multivariate normal populations for a give Type I error rate, successfully discovered the difference between the mean vectors some of the time. The percent of successful discoveries is the power of the test. Likewise, each model failed to successfully reject the null hypothesis some of the time, the percent of which is Type II error. In other words, upon simulating two multivariate normal distributions that have different mean vectors and also meet the assumptions of the particular multivariate model under study, samples can be repeatedly drawn from the two multivariate normal distributions and a multivariate test statistic can be calculated each time. With repeated draws (replications) of the samples from the population, one can obtain the power, that is, one can capture the percent of times that the test statistic from the multivariate model successfully rejects the null hypothesis at a given Type I error rate.

The Monte Carlo simulation just described was used to obtain the power available to each of the three multivariate models being studied for the purpose of comparison. However, to carry out the Monte Carlo simulation, the following independent variables had to be specified: the effect size, the sample size drawn from the multivariate normal populations, the number of dependent variables, the correlation between the posttests and corresponding pretests, as well as the correlation between the unrelated pretest and posttest measurements, the Type I error rate, and the number of dependent variables that are statistically significant. Each of these independent variables is discussed in what follows.

## Independent Variables

### Effect Size between Treatment and Control Groups

Cohen (1988) has used the index *d* to define the difference between the means of two univariate normal populations. Specifically, Cohen defines *d* as the difference between "population means expressed in raw (original measurement) unit" divided by "the standard deviation of either population (since they are assumed equal)" (p. 20). Cohen further defines effect sizes of small, medium, and large to be 0.2, 0.5, and 0.8, respectively. Note that if there were no difference between the means in the univariate normal populations under consideration, then the effect size would be zero.

Cohen (1988) provides a rationale for his values, stating that small, medium, and large effect sizes are relative, yet also useful. He states the following:

> The terms "small," "medium" and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation. . . . In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available. (p. 25)

In this study, Cohen's assessment of usefulness is accepted and his effect sizes of small (0.2), medium (0.5), and large (0.8) have been extended to the multivariate case. These effect sizes are assigned to each statistically significant dependent variable in the dependent vector so that the multivariate distribution representing the treatment group is separated from the control group distribution in a uniform manner. That is, the effect size remains constant across the significant dependent variables in the dependent vector. The population mean vectors that have been selected for use in the Monte Carlo simulation for this dissertation are displayed in Table 1 for different numbers of dependent variables (which is discussed later in greater detail).

Table 1

*Effect Sizes between Treatment and Control Groups for the Population Mean Vectors*

| Effect size | Mean vectors |
|---|---|
| | $p = 2$ |
| $d = .2$ | $\begin{bmatrix} 0.20 \\ 0 \end{bmatrix} \begin{bmatrix} 0.20 \\ 0.20 \end{bmatrix}$ |
| $d = .5$ | $\begin{bmatrix} 0.50 \\ 0 \end{bmatrix} \begin{bmatrix} 0.50 \\ 0.50 \end{bmatrix}$ |
| $d = .8$ | $\begin{bmatrix} 0.80 \\ 0 \end{bmatrix} \begin{bmatrix} 0.80 \\ 0.80 \end{bmatrix}$ |
| | $p = 3$ |
| $d = .2$ | $\begin{bmatrix} 0.20 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.20 \\ 0.20 \\ 0.20 \end{bmatrix}$ |
| $d = .5$ | $\begin{bmatrix} 0.50 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.50 \\ 0.50 \\ 0.50 \end{bmatrix}$ |
| $d = .8$ | $\begin{bmatrix} 0.80 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.80 \\ 0.80 \\ 0.80 \end{bmatrix}$ |
| | $p = 5$ |
| $d = .2$ | $\begin{bmatrix} 0.20 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.20 \\ 0.20 \\ 0.20 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.20 \\ 0.20 \\ 0.20 \\ 0.20 \\ 0.20 \end{bmatrix}$ |

(*table continues*)

Table 1 (*continued*)

| Effect size | Mean vectors |
|---|---|
| | $p = 5$ |
| $d = .5$ | $\begin{bmatrix}0.50\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0.50\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\0.50\end{bmatrix}$ |
| $d = .8$ | $\begin{bmatrix}0.80\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0.80\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0.80\\0.80\\0.80\end{bmatrix}$ |
| | $p = 8$ |
| $d = .2$ | $\begin{bmatrix}0.20\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.20\\0.20\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.20\\0.20\\0.20\\0.20\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.20\\0.20\\0.20\\0.20\\0.20\\0.20\\0\\0\end{bmatrix}\begin{bmatrix}0.20\\0.20\\0.20\\0.20\\0.20\\0.20\\0.20\\0.20\end{bmatrix}$ |
| $d = .5$ | $\begin{bmatrix}0.50\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\0.50\\0.50\\0\\0\end{bmatrix}\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\0.50\\0.50\\0.50\\0.50\end{bmatrix}$ |

(*table continues*)

Table 1 (*continued*)

| Effect size | Mean vectors |
|---|---|
| $d = .8$ | $p = 8$ $\begin{bmatrix}0.80\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0\\0\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0.80\\0.80\\0\\0\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0.80\\0.80\\0.80\\0.80\\0\\0\end{bmatrix}\begin{bmatrix}0.80\\0.80\\0.80\\0.80\\0.80\\0.80\\0.80\\0.80\end{bmatrix}$ |

**Sample Size**

A simulation study by Bonate (2000) used $n_1 = n_2 = 10$ to compare and contrast 11 methods of examining pretest-posttest within the univariate context. Likewise, O'Brien, Parenté, and Schmitt (1982) used $n_1 = n_2 = 10$ to evaluate four common MANOVA criterion tests (Wilks's lambda, Roy's greatest root, Hotelling-Lawley trace, and Pillai's trace) with regard to the robustness of these test statistics under varying levels of bias. Stevens (1980) used 15, 25, 50, and 100 subjects per group to develop an approximating table to determine power in an independent groups design (control compared to treatment). Jamieson (1995) used 25 subjects per group in a computer simulation to examine the effects of a negative correlation between baseline and change on two measures of change, namely, change from baseline scores and

covariance adjusted scores. Tu et al. (2005) used sample sizes of 10, 20, and 30 per group in a Monte Carlo simulation study to examine six different methods typically used for measuring change in a univariate context.

This dissertation expands the sample size dimension to cover the greater number of scenarios that may be encountered in the social sciences and in so doing follows in the steps of two prior Monte Carlo simulations involving multivariate analyses that have used larger sample sizes. In an unpublished dissertation, Heiny (2006) used samples of 50, 100, 200, and 500 subjects per group to examine discriminant analyses as a follow-up to a significant MANOVA. Expanding on Heiny's study, Chandran (2009) used samples of 100, 250, and 500 per group to examine the partial *R*-square and *F* test criteria in stepwise discriminant analysis as a follow-up to a significant MANOVA.

One perspective on the values used by the previously noted authors is achieved if they are considered against a backdrop of the per group sample size requirements to detect a small, medium, and large effect size as defined by Cohen (1988), namely, to detect effect sizes of 0.2, 0.5, and 0.8 standardized difference units between two means of univariate normal populations. Using SAS® PROC POWER and a Type I error rate of 0.05 and power of 80%, an equal variance *t*-test used to carry out a two-sided test for inequality between group means will require 394 subjects per group to detect Cohen's small effect size of 0.2, 86 subjects per group to detect a medium effect size of 0.5 and 26 subjects per group to detect a large effect size of 0.8.

In this study, per group sample sizes of 25, 50, 100, and 250 were used. These values cover the range of sample sizes in the articles cited previously and also correspond to the range of sample sizes required to detect the effect sizes that Cohen (1988) believes capture a majority of the experiments in the social sciences.

**Number of Dependent Variables**

In an unpublished dissertation, Schneider (2002) performed a simulation examining discriminant analysis as a post hoc follow-up procedure to a significant MANOVA. In the process, a wide range of studies performed in the social sciences were examined. Schneider found that $p$ = 2, 5, and 8 provided a good representation of a small, medium, and large number of dependent variables in MANOVA studies, respectively. Heiny (2006) and Chandran (2009) expanded upon the number of dependent variables Schneider had used. Both dissertations used $p$ = 2, 3, 4, 5, 6, 7, and 8. It is the author's belief that, in general, most researchers use a small to medium number of dependent variables. Based on this idea and these previous Monte Carlo simulations, this study used $p$ = 2, 3, 4, 5, and 8 to examine the behavior of the three multivariate designs.

**Within Correlation and Background Correlation of Pretest and Posttest Scores**

Zimmerman and Williams (1982) point out that "correlated errors are probably the rule rather than the exception in pretest–posttest measurements" (p. 153). Indeed, this is the nature of pretest–posttest designs. It normally is expected that an individual's pretest and posttest scores will not be independent from one another. However,

as some authors have pointed out, the degree to which pretest scores and posttest scores are correlated can vary a great deal. Horst, Tallmadge, and Wood (1975) believe that standardized tests can yield correlations between the pretest and posttest as high as 0.80 and 0.90, and Bonate (2000) observed that the average correlation between pretest and posttest is about 0.6 in psychological research and possibly even higher in medical research. Monte Carlo simulations by Yap (1979) used pretest–posttest correlation values of 0.25, 0.50, and 0.75 to evaluate the accuracy of regression models based on within-subject correlation ($\rho_{within}$). Bonate also used correlations of 0.25, 0.50, and 0.75 in Monte Carlo simulations to compare univariate statistical tests in evaluating different pretest–posttest methods, but also included the values of 0, 0.90, and 0.95. In another Monte Carlo simulation study, Tu et al. (2005) used pretest–posttest correlation values of 0.10, 030, 0.50, 0.70, and 0.90 to analyze six different statistical methods for measuring change in univariate randomized controlled trials. The present simulation mimics a combination of the values used by Bonate and Tu et al. to cover the wide spectrum of possible correlations found in pretest-posttest research. Therefore, pretest scores are systematically varied with regard to the degree of association that they exhibit with the corresponding posttest scores to have correlation values of $\rho_{within}$ = 0, 0.10, 0.30, 0.50, 0.70, and 0.90.

The background correlation ($\rho_{background}$) was also manipulated in this study. In the context of this dissertation, the background correlation is the correlation between the $p$ pretests and the correlation between the $p$ posttests. The values that $\rho_{background}$ may assume are 0.10, 0.30, and 0.50. These values have been presented by Cohen

(1988) as small, medium, and large effect sizes, respectively, when the Pearson product-moment correlation (*r*) is used to express the degree of relationship between two variables. As Cohen notes, *r* values of 0.10, 0.30, and 0.50 explain 1%, 9%, and 25%, respectively, of variation in "either variable which may be predicted by (or accounted for, or attributed to) the variance of the other, using a straight-line relationship" (p. 78). These values span a range of commonly found correlations that exist in a variety of research situations. However, Tabachnick and Fidell (2001) point out that the best choice of dependent variables may be ones that are uncorrelated with one another (i.e., independent). Therefore, in addition to the small, medium, and large values already described, a background correlation of 0 is used. Thus, $\rho_{background}$ may assume the values of 0, 0.10, 0.30, and 0.50 in this dissertation.

The correlation structure submitted to the simulation does not look like a typical correlation matrix that one might expect to see. The reason for this difference is because the simulation must—as was just discussed—account for the correlation between all variables at both the pretest and posttest level. One good way to explain the correlation structure is by way of illustration. An example of four correlation structures (one for each assumed $\rho_{background}$ value of 0.0, 0.1, 0.3, and 0.5) can be seen in Table 2. The illustration uses two dependent variables (each having a pretest and a posttest) and a value of $\rho_{within}$ equal to 0.90. The intersections of columns and rows of each correlation matrix represent each variable $X_{ij}$ where i = 1 for the first dependent variable and 2 for the second dependent variable and j = 1 for the pretest and 2 for the posttest. There are, therefore, four columns and four rows in each correlation matrix

for this example. The two samples are drawn from independent multivariate normal populations assuming that the correlation structure for both control and treatment groups is the same.

**Type I Error Rate**

In this study, Type I error rates (levels of α) of 0.01 and 0.05 are used. These values are typically found in the tables of textbooks (e.g., Rosner, 2000) and articles (Hubbard, Bayarri, Berk, & Carlton, 2003) and are, therefore, representative of the alpha values often used in social science research.

**Number of Significantly Different Dependent Variables**

For each dependent vector of a given size ($p$ = 2, 3, 4, 5, or 8 as described previously), the number of dependent variables in the outcome vector with a statistically significant difference between the treatment and control groups is varied. The purpose of this scheme is to determine if the multivariate models differ in their ability to correctly detect multivariate statistical significance between the treatment and control when different numbers of dependent variables in the outcome vector exhibit univariate statistical significance. Due to the magnitude of possibilities over dependent vectors of size 2, 3, 4, 5, and 8, a representative sample of the possible number of univariate statistically significant variables within each dependent vector of a given size was used. The selected numbers of statistically significant dependent variables for outcome vectors of size 2, 3, 4, 5, and 8 are displayed in Table 3.

Table 2

*Population Correlation Matrices*

| Correlation structure |
|---|
| One ($\rho_{background} = 0.0$) |
| $\rho = \begin{bmatrix} 1.00 & 0.90 & 0.00 & 0.00 \\ 0.90 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.90 \\ 0.00 & 0.00 & 0.90 & 1.00 \end{bmatrix}$ |
| Two ($\rho_{background} = 0.1$) |
| $\rho = \begin{bmatrix} 1.00 & 0.90 & 0.10 & 0.10 \\ 0.90 & 1.00 & 0.10 & 0.10 \\ 0.10 & 0.10 & 1.00 & 0.90 \\ 0.10 & 0.10 & 0.90 & 1.00 \end{bmatrix}$ |
| Three ($\rho_{background} = 0.3$) |
| $\rho = \begin{bmatrix} 1.00 & 0.90 & 0.30 & 0.30 \\ 0.90 & 1.00 & 0.30 & 0.30 \\ 0.30 & 0.30 & 1.00 & 0.90 \\ 0.30 & 0.30 & 0.90 & 1.00 \end{bmatrix}$ |
| Four ($\rho_{background} = 0.5$) |
| $\rho = \begin{bmatrix} 1.00 & 0.90 & 0.50 & 0.50 \\ 0.90 & 1.00 & 0.50 & 0.50 \\ 0.50 & 0.50 & 1.00 & 0.90 \\ 0.50 & 0.50 & 0.90 & 1.00 \end{bmatrix}$ |

*Note.* $p = 2$, $\rho_{within} = 0.90$.

Table 3

*Number of Statistically Significant Dependent Variables per Total Dependent Variables*

| | $p_{significant}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| *p* | 1 | 2 | 3 | 4 | 5 | 6 | 8 |
| 2 | x | x | - | - | - | - | - |
| 3 | x | - | x | - | - | - | - |
| 4 | x | x | - | x | - | - | - |
| 5 | x | - | x | - | x | - | - |
| 8 | x | x | - | x | - | x | x |

## Number of Replications

Studies have been done previously that investigated the effects of the number of replications used in Monte Carlo simulations. In an unpublished dissertation, Supawan (2004) examined six published articles that discussed the number of replications necessary for regression simulation studies. The number of replications was increased until the results were stable. It was found that fewer replications were needed to obtain consistent results for power than for Type I error. It was recommended that around 1,250 replications should be used for power, and between 4,200 and 4,600 replications should be used for Type I error.

Preecha (2004) examined the number of replications necessary for power and Type I error in ANOVA simulations in an unpublished dissertation. The conclusions were similar to those of Supawan (2004) in that more replications were required to examine Type I error than power. Preecha recommended that the number of replicates for power be approximately 5,000, and that the number of replicates for Type I error be between 5,000 and 10,000. Based on these two studies, it is reasonable to expect that 5,000 replications will provide stable results when examining power. Thus, 5,000 replicates were used in the simulations presented in this study.

## Test Statistic

O'Brien et al. (1982) evaluated the robustness of four commonly used MANOVA statistics, namely, Wilks's lambda, Roy's largest root test, Hotelling-Lawley trace, and the Pillai-Bartlett trace by systematically altering the level of restricted sampling in the multivariate distributions underlying these tests. Because Wilks's lambda was found to be the least affected when the underlying distributions were restricted, it was used in this study. However, the authors noted that when there are only two groups being compared, all four of the test statistics are equal.

## Procedures

A Monte Carlo simulation was performed using SAS IML (Interactive Matrix Language) and SAS PROC GLM (General Linear Models). Two independent *p*-multivariate normal populations having mean vectors $\mu_1$ and $\mu_2$ were simulated for each of the scenarios previously described. The effect size, *d*, separating the simulated data between the two multivariate normal populations was set by manipulating the

mean values within the mean vector corresponding to each distribution. The effect size describing the difference between $\mu_1$ and $\mu_2$ was set to the specified values that were discussed earlier and the number of significantly different posttest scores ($p_{significant}$) displaying the given effect size was also varied in the manner described above. The same effect size was assumed across all of the statistically significant posttest scores in a given simulation. The correlation matrix, $\rho$, was also constructed with SAS IML using values previously described. The same correlation matrix was used for each of the two populations that were compared within each simulated scenario.

A random sample of size $n$ was drawn from each of the two populations 5,000 times. The number 5,000 was used for the reasons previously discussed. The samples were then evaluated using each of the three multivariate models by SAS PROC GLM at each of the two levels of $\alpha$ (0.01 and 0.05) previously discussed. SAS PROC GLM tested the hypothesis that the two group populations were the same given the effect sizes and correlation matrices that were assumed (i.e., the assumed scenario). A statistically significant test statistic (i.e., Wilks's lambda having a corresponding $p$-value less than or equal to $\alpha$) meant that SAS PROC GLM had successfully detected an a priori difference between the treatment and control groups. The same simulation was performed for each possible scenario using each of the three pretest-posttest designs that were the focus of this dissertation. The power was calculated for each scenario by assessing the percent of detections ($p$-value $\leq 0.01$ or $p$-value $\leq 0.05$) that occurred in the 5,000 replications.

# CHAPTER IV

# ANALYSIS

The following research question is addressed in this section:

Q When pretest and posttest scores are collected, how does statistical power under different sample sizes, effect sizes, numbers of dependent variables, and degrees of correlation within and between the pretest and posttest scores compare between a MANOVA that uses change scores (posttest minus pretest) as dependent variables, a MANOVA that uses only posttest scores as dependent variables and a MANCOVA that uses posttest scores as dependent variables and pretest scores as covariates?

To address this research question, simulated statistical power was calculated for each of the three multivariate methods while systematically varying each of the aforementioned independent variables under study in this dissertation. The resulting number of power estimates for each of the statistical models considered may be determined by multiplying together the number of levels that have been examined for each of the independent variables. There were three levels for effect size (0.2, 0.5, and 0.8), four levels for sample size (25, 50, 100, and 250), five levels for the number of outcome variables (2, 3, 4, 5, and 8), six levels for the within correlation (0, 0.1, 0.3, 0.5, 0.7, and 0.9), four levels for the background correlation (0, 0.1, 0.3, and 0.5), two levels for alpha (0.01 and 0.05), and between two and five levels for the number of significantly different dependent variables (depending on the number of outcome variables), which totals to 15. Thus, 3 x 4 x 6 x 4 x 2 x 15 = 8,640 Monte Carlo

simulations were required to calculate the power for each of the three statistical models, or 25,920 total scenarios. The power for these scenarios was calculated by running each one 5,000 times and calculating the proportion of times that the given method successfully rejected the null hypothesis of no group effect for the given alpha level. In the dataset on the DVD-ROM and the figures that follow, the method with change scores as the outcome variables is denoted by the term, change score, the method with posttests as the outcome variables and pretests as covariates is denoted by the term, MANCOVA, and the method utilizing only posttest scores as outcome variables is denoted by the term, posttest only.

The attached DVD-ROM contains the simulated power for each statistical model and for each unique cross-classification of the independent variables described above. Here, the multivariate method with the greatest statistical power for a given scenario would be the preferred method over the other two. Each of the independent variables can be examined individually in order to understand under what conditions one method might be superior to the other two. Also, one or more combinations of independent variables that give one method an advantage over the other two can be singled out. Researchers confronted with which of these three models to use will obviously benefit from knowing how each independent variable while holding all others constant can affect the models with respect to statistical power, as well as how the independent variables interact with one another to affect statistical power. Therefore, the following two sections take steps to provide a better understanding of

how the independent variables and their interactions relate to statistical power for the three multivariate models under examination.

First, the influence of each of the independent variables on statistical power has been examined across the three multivariate models while controlling for all other independent variables. Here, the average power for each method across the levels of the particular independent variable of interest was examined when the levels of all the other independent variables were collapsed. In other words, the marginal power for each independent variable of interest was examined. For example, the change score method, MANCOVA method, and the posttest only method were compared across each level of $\rho_{within}$ (correlation between pretest and corresponding posttest) by obtaining an average for each $\rho_{within}$ level by combining all levels of all other independent variables. In this fashion, the influence of $\rho_{within}$ on power for each model was isolated and examined.

Second, the circumstances under which one statistical method tended to have superior statistical power relative to the other two were explored. Since the interaction between multiple independent variables could cause one multivariate method to have greater statistical power than the other two, the scenarios that produced greater simulated power in a multivariate method than in the other two were grouped together. The group of scenarios belonging to a given multivariate method—when that method exhibited higher power than the other two—was profiled to determine the underlying characteristics of that particular group of scenarios. In this fashion, a general understanding was formed concerning what combination of independent variable values

would lead to higher statistical power for a given statistical method relative to the other two.

## Relationship Between Each Independent Variable and Statistical Power

Effect size was examined relative to the three multivariate methods while averaging across all of the other independent variables. As expected, the power for each of the methods increased as effect size increased. MANCOVA with posttests as the outcome variables and pretests as covariates displayed greater power than the other two methods across all values of effect size (0.2, 0.5, and 0.8). The change score method had less power than the MANCOVA method at each time point, but greater power than the posttest only method for effect sizes of 0.2 and 0.5. However, at an effect size of 0.8, the posttest only method displayed greater statistical power than the change score method. Figure 1 gives a graphical representation of these findings.

In the same manner, power was compared at each level of sample size for each of the three methods. As expected, power increased for all three multivariate methods as sample size increased. The MANCOVA method consistently displayed greater statistical power than the other two methods as can be seen in Figure 2. The change score method consistently displayed the second best statistical power, and the posttest method displayed the lowest power at each sample size. It may be of interest to note that as the sample sizes increased, the power of the posttest only method approached that of the change score method.

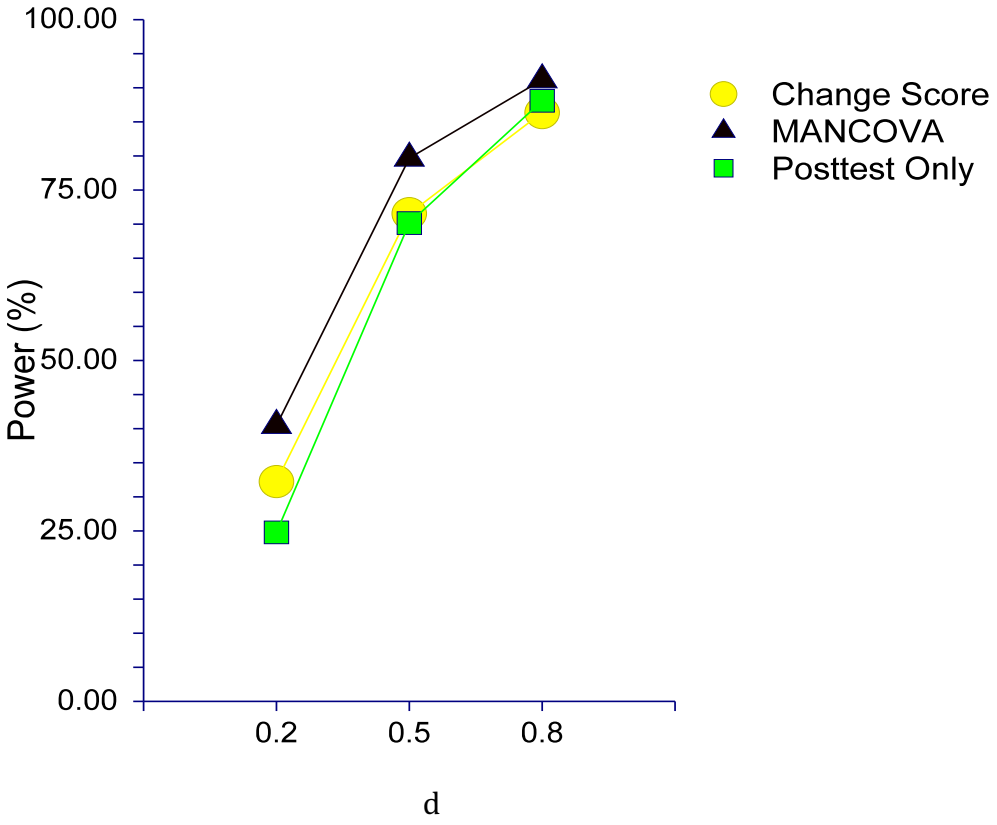


*Figure 1.* Power for effect size, controlling for all other independent variables.

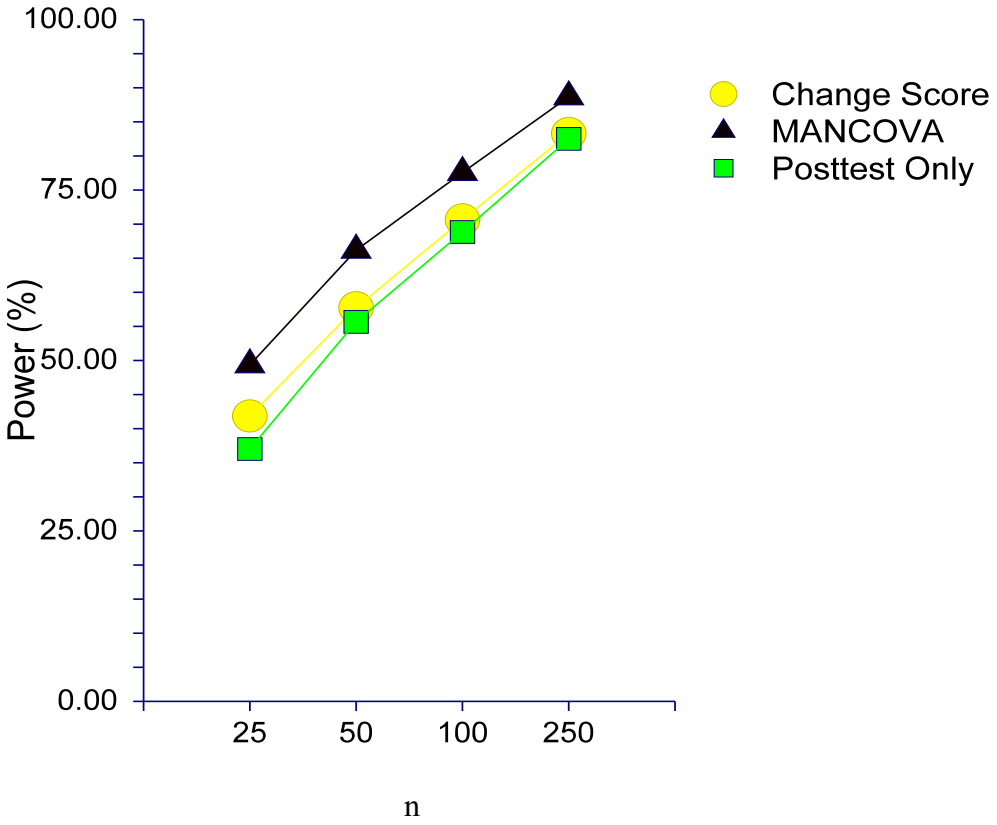


*Figure 2.* Power for sample size, controlling for all other independent variables.

As can be seen in Figure 3, the three multivariate methods are consistent with respect to order of superior statistical power when examined at the two levels of alpha. The MANCOVA method has greater power at both levels of alpha, followed by the change score method. The posttest only method was less powerful than the other two multivariate methods at both levels of alpha. As expected, the statistical power for each method increased as the level of alpha increased.

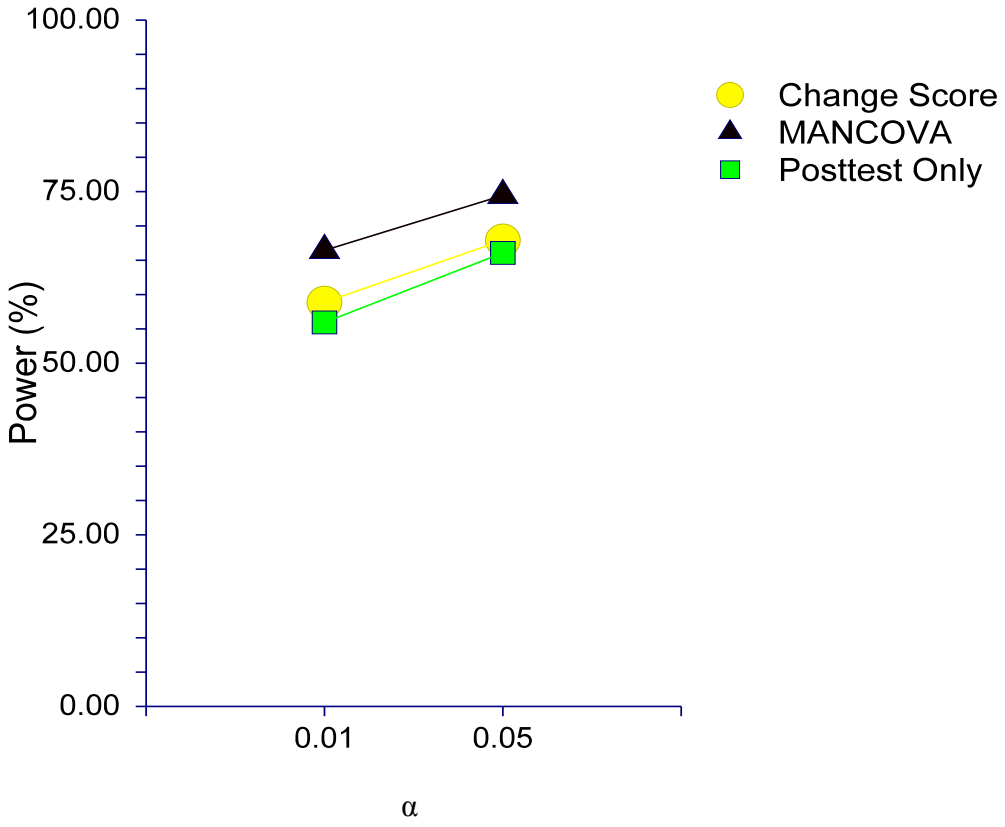


*Figure 3.* Power for levels of significance, controlling for all other independent variables.

Comparing the within correlation (correlation between pretests and their corresponding posttests), it can be seen in Figure 4 that the three multivariate methods behaved differently with regard to statistical power at various levels of correlation. The posttest only method had relatively consistent power across all levels of within correlation, which was to be expected since this method ignores the relationship between the pretests and posttests. It had the second highest power at lower values of within correlation, but because the change score method and MANCOVA method

improved as within correlation increased while the posttest only method remained fairly constant, the posttest only method had the lowest power at higher values of within correlation. The MANCOVA method had the highest statistical power for within correlation values at 0, 0.1, 0.3, 0.5, and 0.7. However, the change score method slightly surpassed the MANCOVA method at $\rho_{within} = 0.9$. This occurred despite the fact that the change score method had the least power at $\rho_{within} = 0$, 0.1, and 0.3.

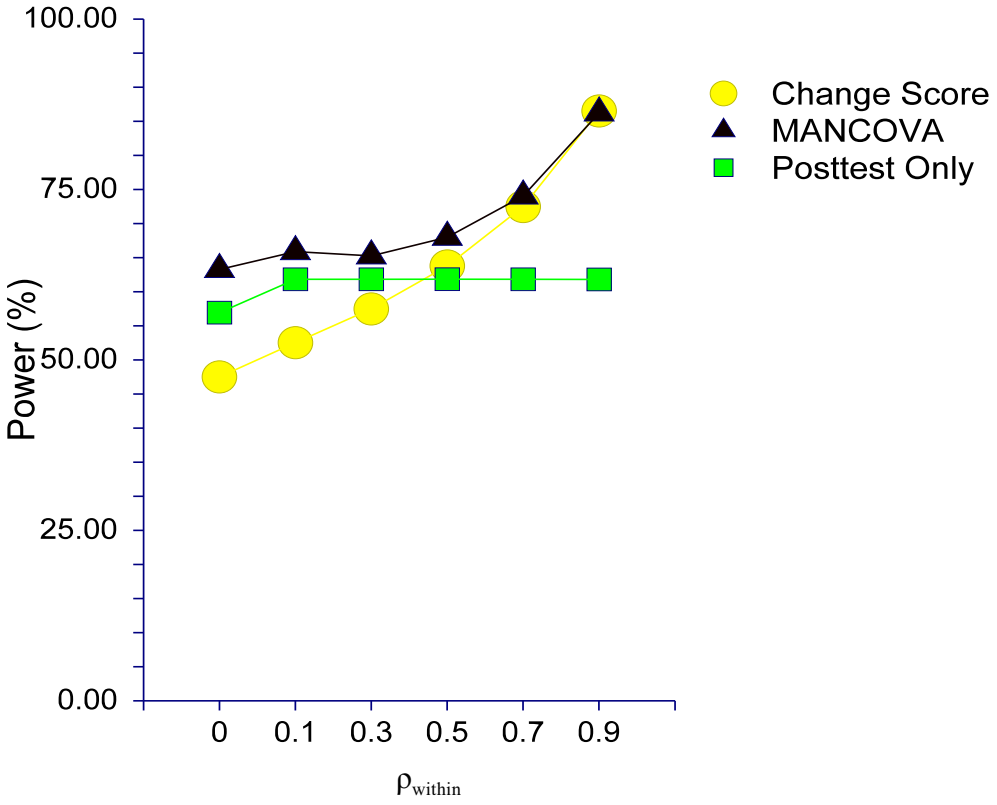


*Figure 4.* Power for within correlation, controlling for all other independent variables.

In Figure 5, it can be seen that the MANCOVA method consistently had greater power than the change score and posttest only methods at each background correlation value, and that its power increased as the background correlation increased. The power of the change score method was nearly identical to that of the posttest only method at $\rho_{background} = 0$, but as the background correlation increased, the change score method was slightly better than the posttest only method. The posttest method displayed a near constant level of power across all of the values, but was consistently lower than the other two methods at $\rho_{background} = 0.1$, 0.3, and 0.5.

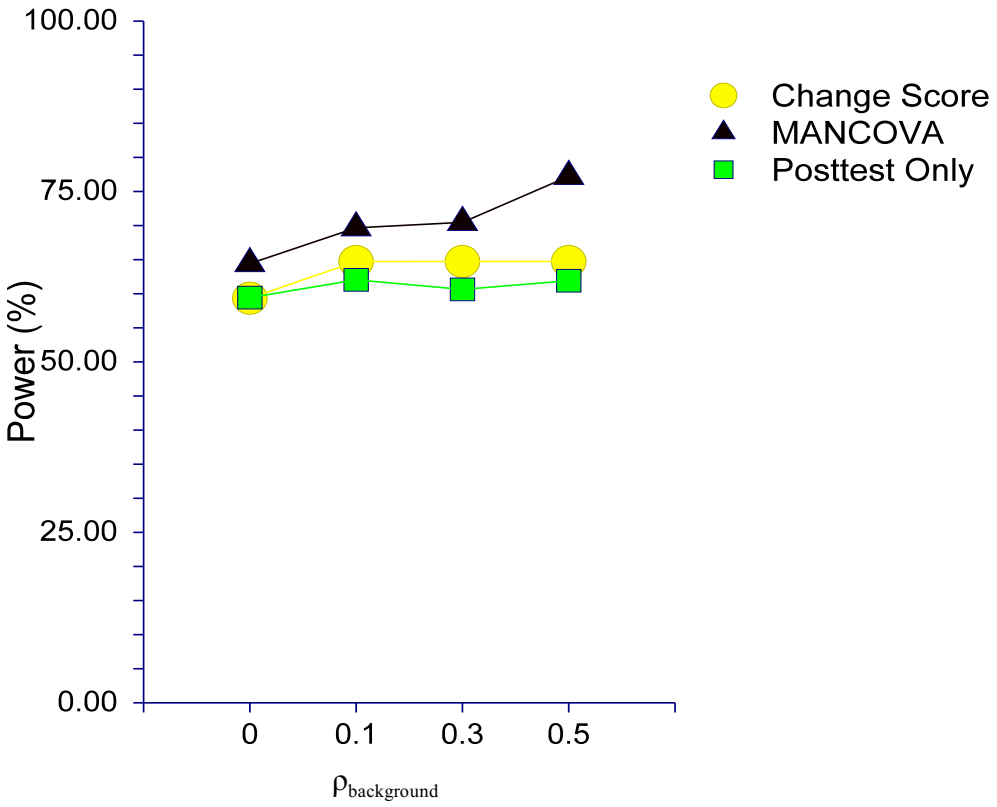


*Figure 5.* Power for background correlation, controlling for all other independent variables.

The three multivariate methods maintained the same ranking across the number of dependent variables when averaging across all other independent variables as can be seen in Figure 6. The MANCOVA method had the highest power, followed by the change score method, while the posttest only method exhibited the lowest power for $p$ = 2, 3, 4, 5, and 8. There appeared to be a drop in power for all three methods at $p$ = 4 as well as a slight drop at $p$ = 8, which could be an artifact of the number of significantly different dependent variables that were chosen for examination in the scenarios with an even number of dependent variables.

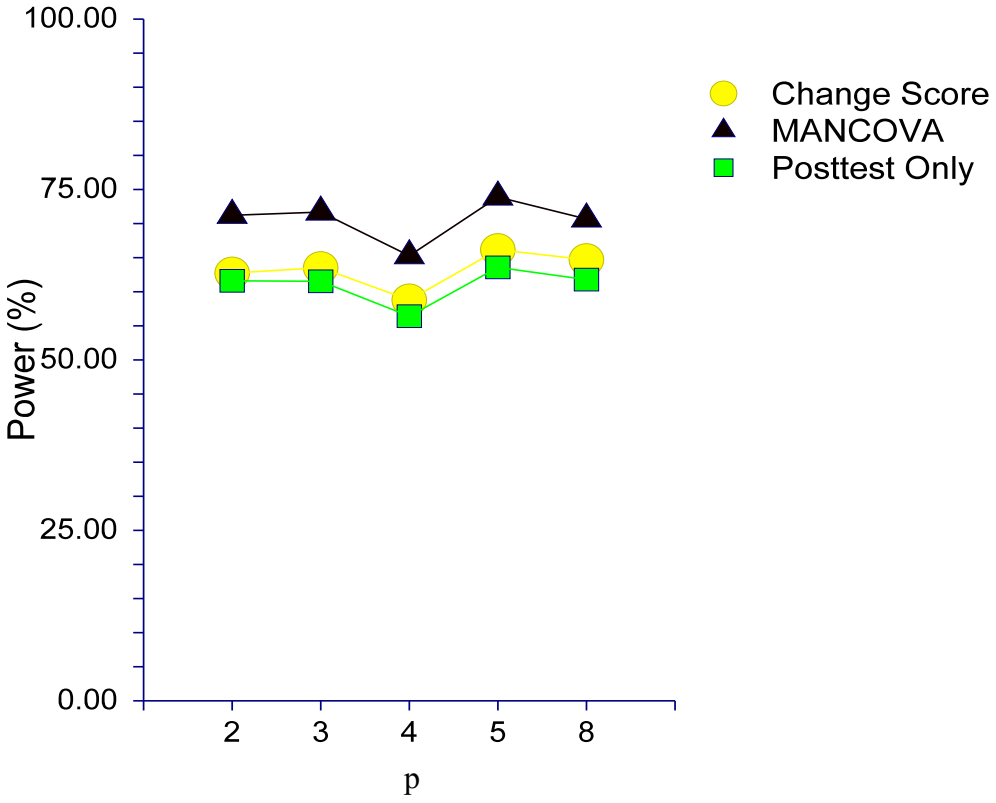


*Figure 6.* Power for the number of dependent variables, controlling for all other independent variables.

Since the number of significantly different dependent variables was unique to each level of number of dependent variables, Figures 7 through 11 display the simulated power for the three multivariate methods by the number of dependent variables. As one would have expected, the simulated power to detect a significant difference between two treatment groups increased for all three methods as the number of significantly different dependent variables, $p_{significant}$, increased. Again, the MANCOVA method had the highest statistical power at each level of significantly different dependent variables within each level of number of dependent variables. The change score method had either the lowest power of the three or was nearly equivalent to the posttest method at $p_{significant} = 1$, regardless of the number of dependent variables. This pattern appears to have held true until more than half of the dependent variables were significantly different, at which point the change score method became more powerful than the posttest method. When all of the dependent variables were significantly different from one another, the simulated power of the change score method approached that of the MANCOVA method. Meanwhile, the simulated power for the posttest only method leveled off as the number of significantly different dependent variables increased, and appeared to even drop off when $p$ was medium to large ($p = 5$ and 8).

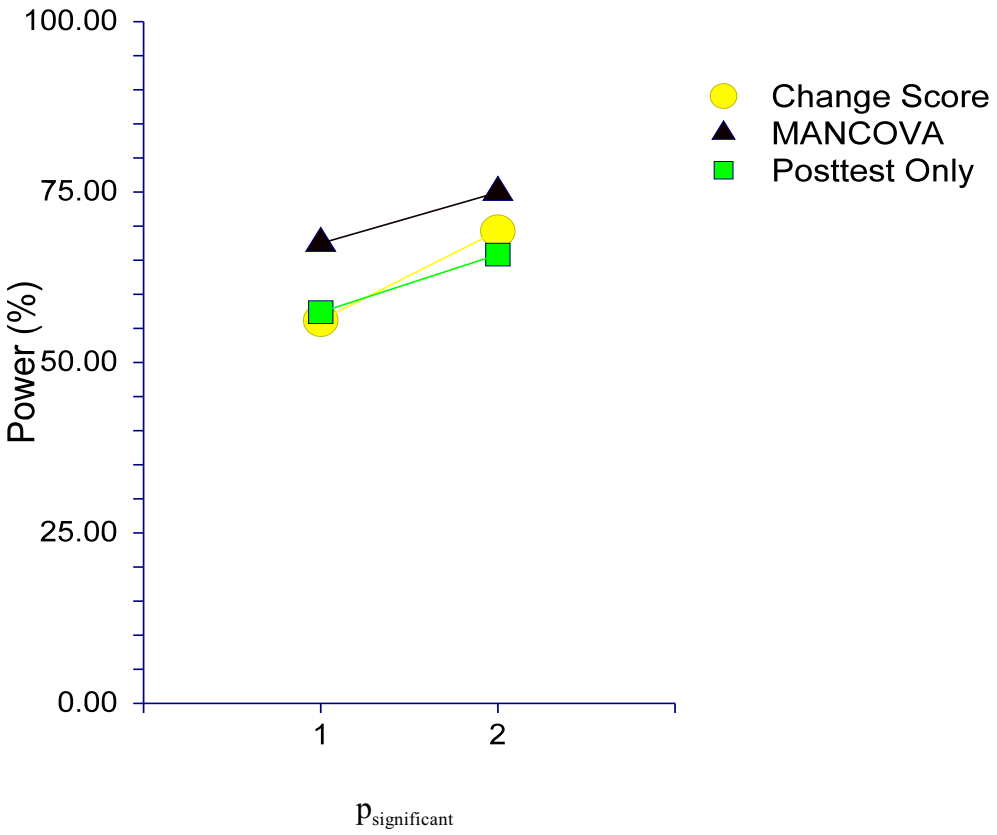


*Figure 7.* Power for the number of significantly different dependent variables, controlling for all other independent variables at $p = 2$.

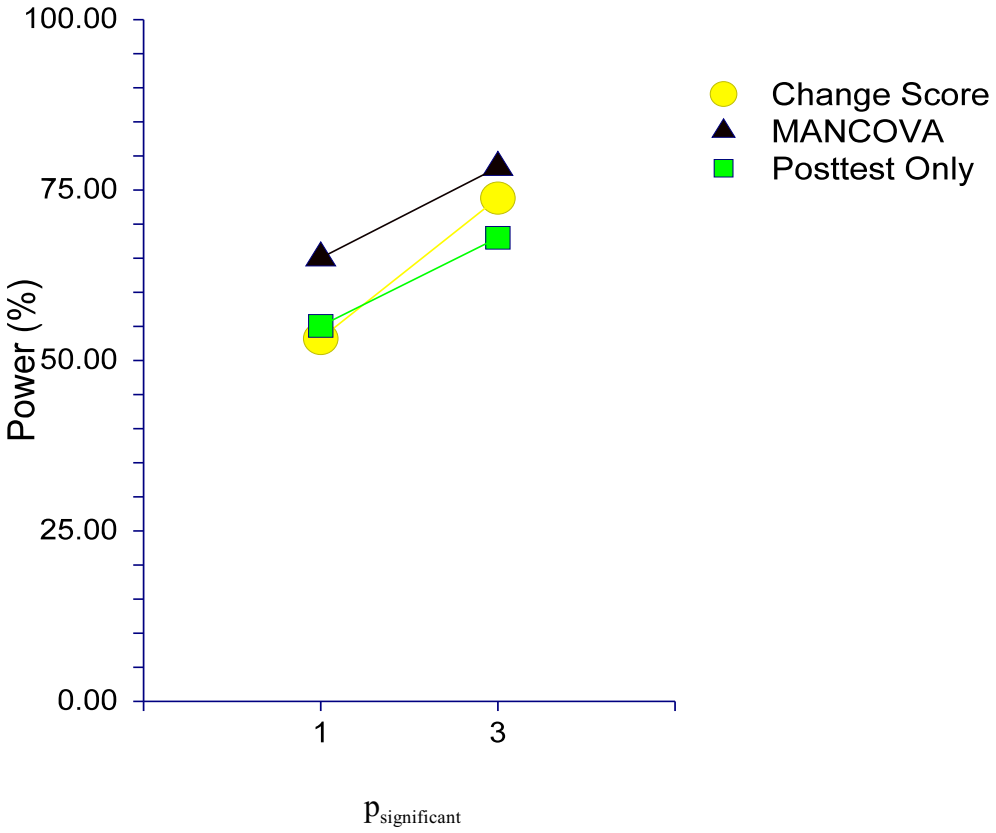


*Figure 8.* Power for the number of significantly different dependent variables, controlling for all other independent variables at $p = 3$.

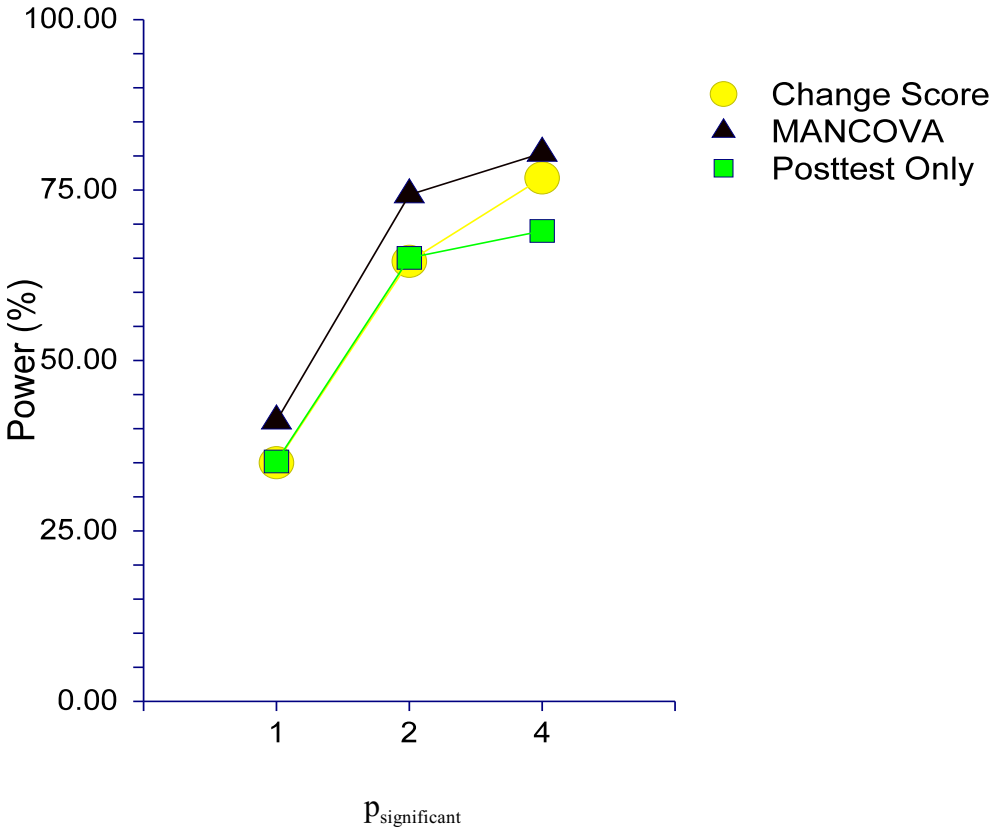


*Figure 9.* Power for the number of significantly different dependent variables, controlling for all other independent variables at $p = 4$.

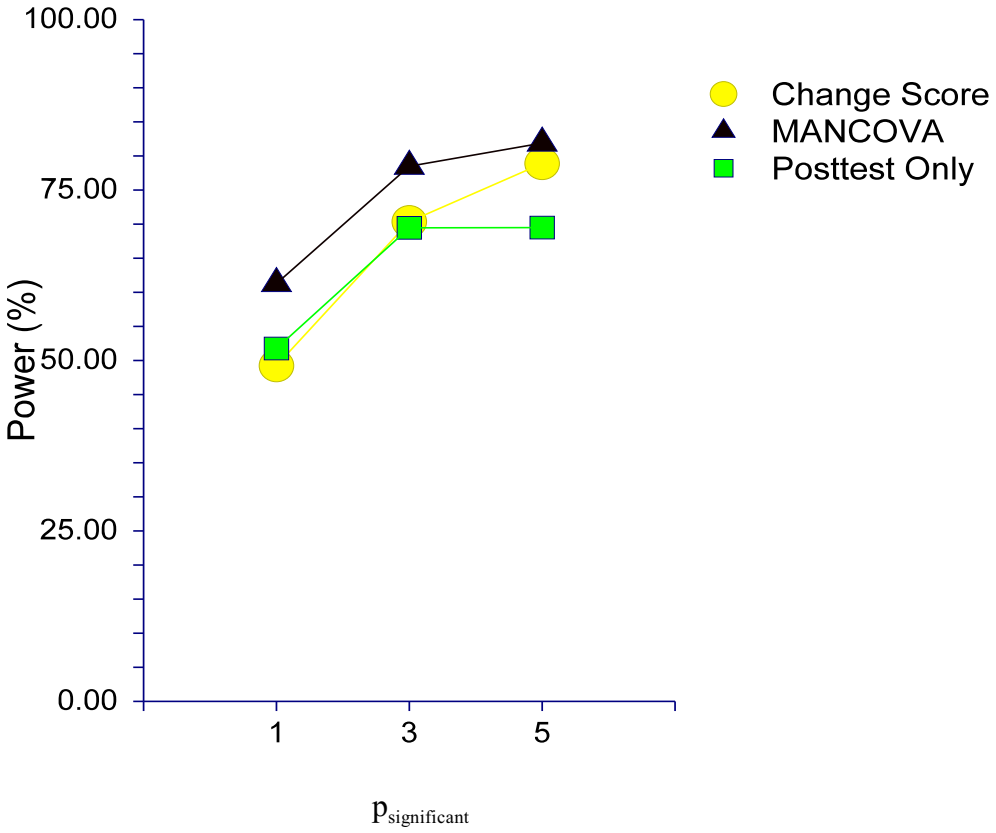


*Figure 10.* Power for the number of significantly different dependent variables, controlling for all other independent variables at $p = 5$.

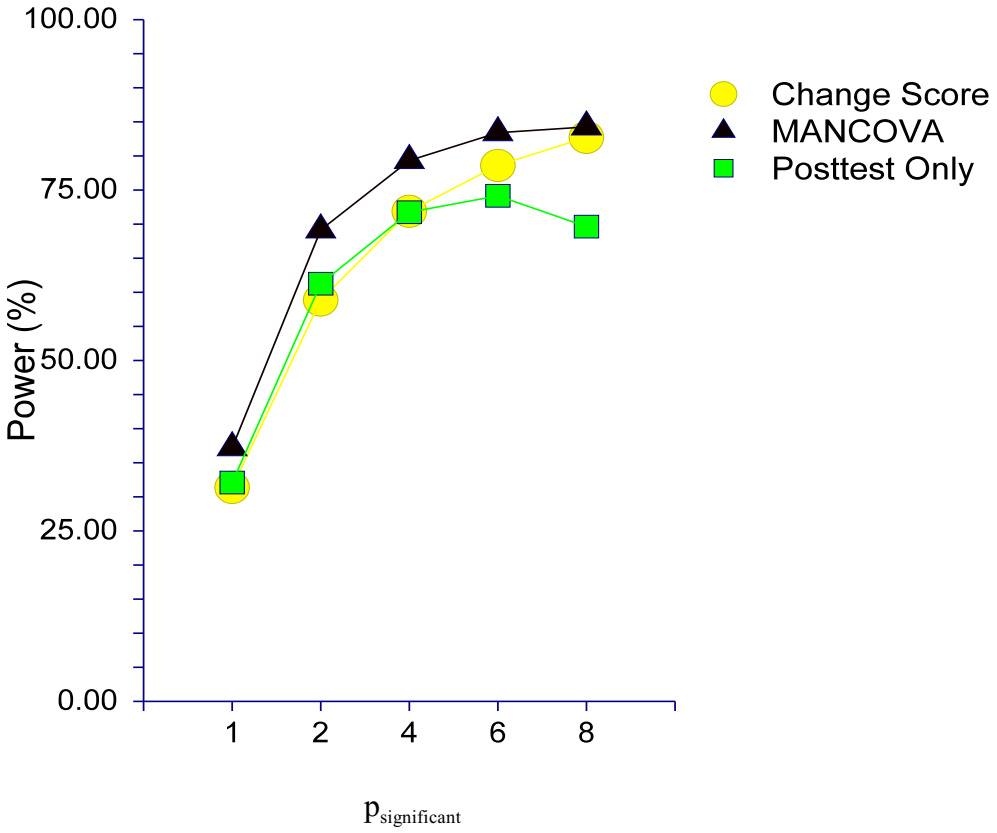


*Figure 11.* Power for the number of significantly different dependent variables, controlling for all other independent variables at $p = 8$.

## Scenarios Where One Model Exhibits Greater Statistical Power Than the Other Two

Although, thus far, it appears that MANCOVA is consistently more powerful than the change score and posttest only methods, this idea does not convey the entire picture. There were a number of scenarios in which the change score or the posttest only methods were superior. There are also scenarios in which one—if not both—methods are equivalent to the MANCOVA method (such as when both the

sample size and effect size are large). However, because the MANCOVA method does seem to be the preferred method, it is appropriate to discuss the performance of the change score and posttest methods relative to it.

A fairly obvious trend can be seen in the dataset on the DVD-ROM when examining the circumstances in which the posttest only method was superior to the other two. The posttest only method appeared to typically have greater statistical power than the MANCOVA method and the change score method when the background correlation was greater than or equal to the within correlation. In other words, if the correlation between pretests and their corresponding posttests was less than the correlation between the unrelated pretests and posttests, the posttest method was often better. However, sample size and effect size also played a part in these results in addition to this interaction between the two types of correlation. As both of these independent variables decreased, the posttest method had more power than the other two methods with greater frequency. Also, the number of dependent variables played a role, since the posttest method had more power with greater frequency as the number of dependent variables increased. At the same time, these results were more prominent when there were fewer significantly different dependent variables. These results did not appear to be dependent upon the level of alpha, as the number of scenarios displaying this phenomenon was nearly equal in each level of alpha. Therefore, the posttest method was typically more powerful when background correlation was greater than within correlation, sample size was low, effect size was low, the number of

dependent variables was large, and the number of significantly different dependent variables was low.

A trend for the scenarios where the change score method had the greatest power compared to the MANCOVA and posttest methods also exists in the dataset included in the DVD-ROM. The change score method displayed greater statistical power than the other two multivariate methods when the correlation between pretests and their corresponding posttests was high, the sample size was small, the number of dependent variables was large, the number of significantly different dependent variables had either a value of one or the highest possible number, and the effect size was small. This phenomenon was evidenced by the proportion of scenarios where the change score method had the highest power increasing as within correlation and the number of dependent variables increased, the sample size decreased, and the number of significantly different dependent variables moved to the minimum or maximum possible values. As was the case when the posttest method was compared to the change score and MANCOVA methods, there was little effect by alpha level. Here, the proportion of scenarios where the change score method was superior does not vary by background correlation either.

# CHAPTER V

# CONCLUSIONS AND DISCUSSION

Chapter IV described the scenarios under which each multivariate method demonstrated greater statistical power than the other two methods. Researchers may find the information afforded by this dissertation useful when planning their studies and subsequent analyses.

## Conclusions

As was expected, statistical power increases for all three of the multivariate models examined as the effect size, sample size, and alpha level increase. Although the MANCOVA method is the most powerful under the majority of instances, there are important circumstances that argue against a one method fits all approach. The difference between all the methods diminishes when effect sizes are high ($d = 0.8$) and the difference between the MANCOVA, and the change score methods dissipates as the within correlation increases and as the number of dependent variables and the number of significantly different dependent variables increases. Additionally, there are important conditions under which the posttest only approach evidences greater power than the other two methods, and this statement is also true for the change score method.

The posttest only method often displayed superior statistical power over the MANCOVA and change score methods when the within correlation was less than or equal to the background correlation. This result occurred with increased frequency if the sample size was small, the number of dependent variables was large, and the number of significantly different dependent variables was small. Therefore, researchers may find they require a smaller sample size and/or are be less likely to commit a Type II error if their study has these characteristics and they use the posttest only method to perform the multivariate analysis. However, it should be pointed out that the scenarios just described may only occur very rarely, if at all. If the correlation between the pretests and corresponding posttests in a study is less than the background correlation between unrelated pretests and posttests, then the researcher would have to question whether or not the pretests and posttests that have been selected are appropriate for use in the study.

The change score method had more statistical power than the other two methods when within correlation was high, the sample size was small, the number of dependent variables was large, the number of significantly different dependent variables was either one or the highest possible number, and the effect size was small. Therefore, researchers would benefit from using the change score method when their studies have these characteristics. A possible explanation for this finding is that the change score method uses information contributed jointly by the pretests and posttests, whereas the posttest method uses less information because the pretest is deleted and the MANCOVA method requires a greater number of degrees of freedom to estimate

parameters. This explanation was, in fact, anticipated by Maxwell and Howard (1981). From the current research it is interesting to note that as sample size increases, the degrees of freedom used by the covariates in the MANCOVA method appear to become relatively less important. Thus, the MANCOVA method is associated with disproportionately increasing statistical power relative to the change score approach as sample size increases.

The above considerations aside, it remains true that the MANCOVA method exhibited greater power under many more scenarios than did either the change score or the posttest only methods. Excluding those scenarios just described and scenarios where the MANCOVA method and one or both of the other methods had 100% power, the MANCOVA method was superior to the other two methods in all other scenarios with regard to power. In the scenarios where the MANCOVA method did not have the highest power, it had less than a 5% difference in power relative to the superior method 98.6% of the time. Therefore, if a researcher must choose a method *a priori* without knowing the characteristics of the study, it is recommended that the researcher use the MANCOVA method. However, if the researcher suspects scenarios compatible with greater power for the posttest only method or the change score method, this dissertation provides a defensible rationale for selecting one of these two alternative methods.

Finally, it is important to remember that power can be adjusted by manipulating factors other than the selected statistical model for analysis. In this dissertation, the implicit assumption has been that if sample size and alpha are held constant then a

researcher would use, for a given scenario defined by the independent variables examined here, the model exhibiting the most power in order to perform his or her analysis. This situation, however, is not always true. At times, one analysis method may be preferred over the other for reasons not directly related to statistical power. For example, the cost of collecting a pretest score might be greater than the cost of increasing sample size. In this case, the posttest only method might be preferred regardless of the scenario, and the lesser amount of power relative to the other methods might be compensated by increasing the sample size. A final example relates to an earlier point made by Fitzmaurice et al. (2004). These authors pointed out that the analysis method selected must directly address the research question at hand. In the multivariate setting, the change score method tests whether there is a statistically significant difference between one or more mean change scores of two or more groups, regardless of whether the baseline values are equal between treatments. On the other hand, the MANCOVA method tests whether one or more mean posttest scores differ significantly between two or more groups after adjusting for differences that may have existed between the pretest scores. The research questions are quite similar but indeed different, and depending on the purpose of the study, one may be preferred over the other. If the change score method answers the research question but exhibits less power for the expected scenario, then perhaps parameters impacting statistical power, such as sample size, should be altered to allow the use of the model that answers the research question best. This dissertation offers insight not only into which model provides the greatest statistical power under a fixed scenario, but also insight into the

degree of power that might need to be compensated for if a multivariate method with less power were used for reasons like those just presented.

## Discussion

This study focused on three methods that could be used to analyze change when multiple pretests and corresponding posttests exist. Two of these methods involve models that utilize both the pretest and posttest scores (MANCOVA with pretest scores as covariates and MANOVA with change scores). The third method ignores pretest information altogether (MANOVA with posttest scores only). While the findings presented above are of considerable applied relevance, this study also serves as a starting point for additional research that examines issues surrounding the analysis of multivariate change.

This dissertation concentrated on situations analyzing the difference in change from pretest to posttest between two independent groups. Certainly, researchers sometimes desire to know the difference in change from pretest to posttest between three or more groups. The work presented here could be extended to cover studies with more than two comparison groups and a comparison of statistical power of the multivariate methods could be performed based on this potential independent variable.

In addition to studying the statistical power for the three multivariate models, Type I error could also be examined. The decision to use one of the three methods examined here instead of the other two should not depend solely on statistical power, but also on how well each method controls for Type I error. If one method displays more power than the others, but differences exist in the nominal alpha level as a

function of sample size, effect size, within correlation, background correlation, and/or the number of dependent variables, then this too constitutes an important area of future investigation.

The multivariate methods surveyed in this dissertation are not the only ones that can be used to analyze multivariate studies where subjects have an array of posttest outcomes with corresponding pretest values available. Bonate (2000) and Tu et al. (2005) present numerous other univariate methods for analyzing situations involving a single pretest and corresponding posttest. These methods could be generalized to the multivariate case and compared with the models examined in this dissertation or with each other. These comparisons could be conducted with respect to power and/or Type I error. A few examples of these methods are using percent change scores, using log-transformed change scores, using ranked normal pretest scores and ranked normal posttest scores, and using both pretest and posttest scores as outcome variables in a multivariate analysis of variance.

While this dissertation focused on scenarios where the pretests of each of the two comparison groups were assumed to be equal, other scenarios commonly occur in research. In the univariate case, Bonate (2000) provided a number of different possible scenarios and performed Monte Carlo simulations to address them. Such scenarios could easily be extended to the multivariate realm and examined using Monte Carlo simulations similar to the ones used in this dissertation. Some of the scenarios presented by Bonate are when subjects are put into groups based on their pretest scores, when the variance of posttest scores does not equal the variance of

pretest scores, or when marginal distributions of the pretest scores and posttest scores are not normally distributed. These issues, with respect to both power and Type I error, provide a number of important areas for future research.

REFERENCES


Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93-114.

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society, 35*, 180-185.

Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society, 9*(2), 176-197.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York, NY: McGraw-Hill.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs.* Boca Raton, FL: CRC Press.

Burgess, E. W., Locke, H. J., & Thomes, M. M. (1971). *The family.* New York, NY: Van Nostrand Reinhold.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Hopewell, NJ: Houghton Mifflin.

Chandran, R. K. (2009). *The effectiveness of stepwise discriminant analysis as a follow up procedure to a significant MANOVA using both the F-statistic and R-square criterion* (Unpublished doctoral dissertation). University of Northern Colorado, Greeley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Lawrence Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online, 9*(1), 37-55.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74*(1), 68-80.

Delaney, H. D., & Maxwell, S. E. (1980). The use of covariance in tests of attribute-by-treatment interactions. *Journal of Educational Statistics, 5*, 115-128.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Societry of London Series A, 222,* 309-368.

Fitzmaurice, G. (2001). A conundrum in the analysis of change. *Nutrition, 17*(4), 360-361.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley-Interscience.

Gottman, J. M., & Krokoff, L. J. (1989). Marital interaction and satisfaction: A longitudinal view. *Journal of Consulting and Clinical Psychology, 57*(1), 47-52.

Gottman, J. M., & Krokoff, L. J. (1990). Complex statistics are not always clearer than simple statistics: A reply to Woody and Costanzo. *Journal of Consulting and Clinical Psychology, 58*(4), 502-505.

Haase, R. F., & Ellis, M. V. (1987). Multivariate analysis of variance. *Journal of Counseling Psychology, 34*(4), 404-413.

Heiny, E. L. (2006). *The effectiveness of stepwise discriminant analysis as a post hoc procedure to a significant MANOVA* (Unpublished doctoral dissertation). University of Northern Colorado, Greeley.

Hershberger, S. L. (2005). History of multivariate analysis of variance. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 864-869). New York, NY: Wiley.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. (1975). *A practical guide to measuring project impact on student achievement.* Washington, DC: U.S. Government Printing Office.

Hubbard, R., Bayarri, M. J., Berk, K. N., & Carlton, M. A. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician, 57*(3), 171-182.

Huck, S. W., & McLean, R. A. (1975, July). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*(4), 511-518.

Jamieson, J. (1995). Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement, 55*(1), 38-46.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Kessler, R. C. (1977). The use of change scores as criteria in longitudinal survey research. *Quality and Quantity, 11*, 43-66.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47,* 121-150.

Locke, H. J., & Wallace, K. M. (1959). Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living, 21,* 251-255.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*(4), 421-437.

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement, 18*(3), 437-451.

Lord, F. M. (1960, June). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*(290), 307-321.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). Madison, WI: University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304-305.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*(3), 309-327.

Markus, G. (1980). *Models for the analysis of panel data.* Beverly Hills, CA: Sage.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). New York, NY: Psychology Press.

Maxwell, S. E., & Howard, G. S. (1981). Change scores—necessarily anathema? *Educational and Psychological Measrement, 41,* 747-756.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement, 18*, 47-55.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review, 25,* 3-28.

O'Brien, P. N., Parenté, F. J., & Schmitt, C. J. (1982). A Monte Carlo study on the robustness of four MANOVA criterion tests. *Journal of Statistical Computation and Simulation, 15,* 183-192.

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42*(1), 73-97.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*(1), 85-86.

Preecha, C. (2004). *Numbers of replications required in ANOVA simulation studies* (Unpublished doctoral dissertation). University of Northern Colorado, Greeley.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* New York, NY: Wiley.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Cambell, W. Meredth, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171-209). New York, NY: Springer.

Rosner, B. A. (2000). *Fundamentals of biostatistics* (5th ed.). Pacific Grove, CA: Duxbury.

Roy, S. N. (1946). Multivariate analysis of variance: the sampling distribution of the numerically largest of the p-statistics on the non-null hypotheses. *Sankhya, 8*(Pt. 1), 15-52.

Schneider, M. K. (2002). *A Monte Carlo investigation of the Type I error and power associated with descriptive discriminant analysis as a MANOVA post hoc procedure* (Unpublished doctoral dissertation). University of Northern Colorado, Greeley.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin, 88*(3), 728-737.

Supawan, P. (2004). *An examination of the number of replications required in regression simulation studies* (Unpublished doctoral dissertation). University of Northern Colorado, Greeley.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Tu, Y. K., Blance, A., Clerehugh, V., & Gilthorpe, M. S. (2005). Statistical power for analyses of changes in randomized controlled trials. *Journal of Dental Research, 84,* 283-287.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24*(3-4), 471-494.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*(1), 59-69.

Woody, E. Z., & Costanzo, P. R. (1990). Does marital agony precede marital ecstasy? A comment on Gottman and Krokoff's "Marital interaction and satisfaction: A longitudinal view." *Journal of Consulting and Clinical Psychology, 58*(4), 499-501.

Yap, K. O. (1979, April). *Pretest-posttest correlation and regression models.* Paper presented at the 63rd annual meeting of the American Educational Research Association (pp. 3-30), San Francisco, CA. (ERIC Document Reproduction Service No. ED174641)

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149-154.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 269-304). Greenwich, CT: JAI Press.

# APPENDIX

## SAMPLE SAS® CODE FOR MONTE CARLO SIMULATIONS

```
*********************************************
**** Justin L. Rogers                       ****
**** SAS code for Monte Carlo simulations ****
**** p=2                                    ****
*********************************************;

dm output 'clear'; dm log 'clear';

options nonotes nodate;

**** Specify location to store permanent datasets ****;
libname dds "C:\Dissertation Datasets";

**** Print start time to log ****;
data _null_;
  start=datetime();
  format start datetime.;
  put start=;
run;

**** Create base dataset to append to ****;
  data dds.p2_alldata_25;
    input group rep effect_size sig_diff within_corr backg_corr alpha
          col1 col2 col3 col4 diff1 diff2;
    cards;
  run;

  data dds.p2_alldata_50;
    input group rep effect_size sig_diff within_corr backg_corr alpha
          col1 col2 col3 col4 diff1 diff2;
    cards;
  run;

  data dds.p2_alldata_100;
    input group rep effect_size sig_diff within_corr backg_corr alpha
          col1 col2 col3 col4 diff1 diff2;
    cards;
  run;

  data dds.p2_alldata_250;
    input group rep effect_size sig_diff within_corr backg_corr alpha
          col1 col2 col3 col4 diff1 diff2;
    cards;
  run;

**** Begin macro to generate data for analyses ****;
%macro mkdata(n=);

  **** Effect Sizes ****;
  %do e=2 %to 8 %by 3;
    data _null_;
      temp="&e";
      e_s=round(temp/10,.1);
      call symput('e_s',e_s);
    run;

    %let e_s=&e_s;
```

```
**** Number of Significantly Different Posttests ****;
%do dv=1 %to 2;

  **** Within Correlation ****;
  %do w=1 %to 11 %by 2;
    data _null_;
      temp="&w";
      w_c1=round(temp/10,.1);
      if w_c1=1.1 then w_c=0;
      else w_c=w_c1;
      call symput('w_c',w_c);
    run;

    %let w_c=&w_c;

    **** Background Correlation ****;
    %do b=1 %to 7 %by 2;
      data _null_;
        temp="&b";
        b_c1=round(temp/10,.1);
       if b_c1=.7 then b_c=0;
       else b_c=b_c1;
       call symput('b_c',b_c);
      run;

      %let b_c=&b_c;

      **** Alpha Level ****;
      %do a=1 %to 5 %by 4;
        data _null_;
          temp="&a";
          alpha=round(temp/100,.01);
          call symput('alpha',alpha);
        run;

        %let alpha=&alpha;

          **** Number of replications per scenario ****;
          %do rep=1 %to 5000;

         **** Create Correlation Matrix ****;
          proc iml;
            R={1    &w_c &b_c &b_c,
               &w_c 1    &b_c &b_c,
               &b_c &b_c 1    &w_c,
               &b_c &b_c &w_c 1   };

           **** Define vector of standard devations ****;
           Ds=Diag({1 1 1 1});

           **** Compute covariance matrix ****;
            S=Ds*R*Ds;

            **** Compute Choleski Root for transformation ****;
            T=Root(S);
```

```
**** Specify number of observations per sample ****;
n=&n;

**** Create GROUP 1 (Treatment) ****;
**** Specify random number seed ****;
Seed1=0;

**** Create data vector using seed ****;
X1=J(n,NRow(S),Seed1);

**** Generate independent normal distribution ****;
X1=rannor(X1);

**** Transform for covariance structure ****;
Y1=X1*T;

**** Create dataset of GROUP1 (Treatment) data ****;
create group1 from Y1;
append from Y1;
close group1;

**** Create GROUP 2 (Control) ****;
**** Specify random number seed ****;
Seed2=0;

**** Create data vector using seed ****;
X2=J(n,NRow(S),Seed2);

**** Generate independent normal distribution ****;
X2=rannor(X2);

**** Transform for covariance structure ****;
Y2=X2*T;

**** Create dataset of GROUP2 (Control) data ****;
create group2 from Y2;
append from Y2;
close group2;

**** End IML ****;
quit;

**** Give the posttest of GROUP1 (Treatment) the ****
**** specified effect size                       ****;
data group1; set group1;
  if &dv=2 then do;
    array col col1-col4;
      do i=2 to 4 by 2;
        col[i]=col[i]+&e_s;
      end;
    drop i;
  end;
    else if &dv=1 then do;
      col2=col2+&e_s;
    end;
    group=1;
  run;
```

```
                    data group2; set group2;
                      group=2;
                    run;

                    **** Merge GROUP1 (Treatment) and GROUP2     ****
                    **** (Control) into one dataset for analyses ****;
                    data allgroups;
                      merge group1 group2;
                      by group;
                      rep=&rep;
                      effect_size=&e_s;
                      sig_diff=&dv;
                      within_corr=&w_c;
                      backg_corr=&b_c;
                      alpha=&alpha;

                      **** create change scores for each pretest and****
                      **** corresponding posttest                  ****;
                      diff1=col2-col1;
                      diff2=col4-col3;

                      keep group rep effect_size sig_diff within_corr
                           backg_corr alpha col1 col2 col3 col4 diff1
                           diff2;
                     run;

                     **** Compile the datasets into one so that there****
                     **** will only be one dataset per sample size  ****;
                     proc append base=dds.p2_alldata_&n data=allgroups
                                 force;
                     run;

  **** Close DO loops ****;
             %end;
           %end;
         %end;
       %end;
     %end;
  %end;

  **** Sort data for BY variable analyses ****;
  proc sort data=dds.p2_alldata_&n;
    by effect_size sig_diff within_corr backg_corr alpha rep;
  run;

**** End macro ****;
%mend;

**** Call macro for each given sample size ****;
%mkdata(n=25);
%mkdata(n=50);
%mkdata(n=100);
%mkdata(n=250);

**** Suppress output ****;
ods listing close;
```

```
**** Begin macro to run the three multivariate analyses on the ****
**** data                                                       ****;
%macro analyze(n=);

  **** BEGINNING OF ANALYSES ****;

  ************************
  **** MANCOVA method ****
  ************************;
  proc glm data=dds.p2_alldata_&n;
    by effect_size sig_diff within_corr backg_corr alpha rep;
    class group;
    model col2 col4=group col1 col3;
    manova h=_all_;
    ods output multstat=p2_mancova_&n;
  run;
  quit;

  **** Select results testing group effect using Wilks Lambda ****;
  **** Output permanent results dataset ****;
  data p2_mancova_&n; set p2_mancova_&n;
    length method $ 20;
    if hypothesis="group";
    if statistic="Wilks' Lambda";

    **** Determine if a Type II Error was committed ****;
    if probf > alpha then type2error=1;
    else type2error=0;
    method="MANCOVA";
  run;

  ****************************
  **** Change Score Method ****
  ****************************;
  proc glm data=dds.p2_alldata_&n;
    by effect_size sig_diff within_corr backg_corr alpha rep;
    class group;
    model diff1 diff2=group;
    manova h=_all_;
    ods output multstat=p2_diff_&n;
  run;
  quit;

  **** Select results testing group effect using Wilks Lambda ****;
  **** Output permanent results dataset ****;
  data p2_diff_&n; set p2_diff_&n;
    length method $ 20;
    if hypothesis="group";
    if statistic="Wilks' Lambda";

    **** Determine if a Type II Error was committed ****;
    if probf > alpha then type2error=1;
    else type2error=0;
    method="Change Score";
  run;
```

```
  ****************************
  **** Posttest Only Method ****
  ****************************;
  proc glm data=dds.p2_alldata_&n;
    by effect_size sig_diff within_corr backg_corr alpha rep;
    class group;
    model col2 col4=group;
    manova h=_all_;
    ods output multstat=p2_post_&n;
  run;
  quit;

  **** Select results testing group effect using Wilks Lambda ****;
  **** Output permanent results dataset ****;
  data p2_post_&n; set p2_post_&n;
    length method $ 20;
    if hypothesis="group";
    if statistic="Wilks' Lambda";

    **** Determine if a Type II Error was committed ****;
    if probf > alpha then type2error=1;
    else type2error=0;
    method="Posttest Only";
  run;

  **** END OF ANALYSES ****;

  **** Merge all results for given sample size together ****;
  data p2_all_res_&n;
    merge p2_mancova_&n p2_diff_&n p2_post_&n;
     by method effect_size sig_diff within_corr backg_corr alpha rep;
     n=&n;
  run;

  **** Sort dataset for calculation of power ****;
  proc sort data=p2_all_res_&n;
   by method n effect_size sig_diff within_corr backg_corr alpha rep;
  run;

**** End macro ****;
%mend;

**** Call macro for each given sample size ****;
%analyze(n=25);
%analyze(n=50);
%analyze(n=100);
%analyze(n=250);

**** Combine all results ****;
data dds.p2_all_res;
  merge p2_all_res_25 p2_all_res_50 p2_all_res_100 p2_all_res_250;
  by method n effect_size sig_diff within_corr backg_corr alpha rep;
run;

**** Calculate proportion of analyses where Type II Error was ****
**** not committed                                            ****;
**** This proportion will be the simulated statistical power ****;
```

```
proc freq data=dds.p2_all_res;
  by method n effect_size sig_diff within_corr backg_corr alpha;
  table type2error / out=p2_power;
run;

**** Allow output ****;
ods listing;

**** Create permanent dataset containing calculated statistical power
****;
data dds.p2_power; set p2_power;
  if type2error=0;
  p=2;
run;

**** Print end time to log ****;
data _null_;
  end=datetime();
  format end datetime.;
  put end=;
run;

**** END OF SAS PROGRAM ****;
```