Simulated Power Study of ANCOVA vs. Repeated Measure Analyses for Two-Way Designs with one Repeated Measure

Julien R. Lemay

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements

for the Ph. D. degree in experimental psychology

School of Psychology

Department of Social Sciences

University of Ottawa

Abstract

Whether one should use an analysis of covariance or a form of difference score test (difference as an outcome or repeated measure) is not always clear. The literature on the topic focused for a while on Lord's paradox which lead to the conclusion that both analyses were equally valid when there is true random assignment. Yet, the issue of which analysis is best was little explored. In an attempt to create a unifying simulation framework that will allow for comparable results when exploring various data structure variations, I will tackle 5 such manipulations by exploring the impact of varying effect size and relationship between time points, violating the homogeneity of the regression slopes assumption, exploring the effect of large systematic baseline differences, the impact of data missing at random, as well as comparing the sample size requirements for a given test power. The programs provided, which allow for tens of millions of simulations to be run in a reasonable time frame (within a day) also puts to rest any ambiguity on the stability of the results. By analyzing Type I error rate and statistical power, I establish that ANCOVA respects the type-I error rate of alpha, and has more power than repeated measure analysis in most cases, but should be avoided when there is a baseline imbalance. Hence, in cases where ANCOVA is applicable, it is preferable to use it over other difference score tests.

Table of Contents

Introduction

There is much debate about the use of Analysis of Covariance (ANCOVA) versus the use of difference scores in pre-post designs. As shown in Kutner, Nachtsheim, Neter, and Li (2005), the ANCOVA model is equivalent to a repeated measure ANOVA (RMANOVA) when the unstandardized regression coefficient between **Time 1** and **Time 2** approaches unity (unity being when there is a perfect relationship between the two scores). In practical terms, the relationship is never perfect, and thus, the question is which of the two equations is best. This debate has taken several forms, but there are still some avenues left to explore. Some arguments were problematic; I loosely classified them in three categories.

First, over the years, many valid concerns have been brought forward, but are often substantiated with only a single hypothetical dataset. One of the limits of this approach is that although the demonstrations are sound, it could very well be a case of Type I or Type II error. In subsequent studies, I used 1 million replications for each parameter variation.

Second, it is quite common for a single method to be studied in isolation. This does not provide differential information with which to compare the methods. Sometimes a concern brought forward in regards to a method can be well founded, but that method could still be relatively superior to the alternatives. Herein I always compare ANCOVA and RMANOVA on the same tests.

Lastly, many statements are made in a very specific context, and get generalized. These generalized statement then get cited, detached from their context. To the credit of those who have fallen into this trap, it usually happened earlier in the development of the literature when less research on specific questions was available. It is the perpetuation of the generalizations of those statements that are problematic at this point, but they are so far removed from their source decades ago, that it is sometimes difficult to discern the source of these misunderstandings. To avoid overgeneralizations, I inspect a wide range of **scenarios**.

My main objectives are as follow. The first is to replicate past findings using a single frame of reference, this will make results easier to compare. I also aim to fill gaps by exploring the manipulated parameters in a more granular and comprehensive fashion, as well as following up on perceived weaknesses of the ANCOVA in the face of a violation of the homogeneity of regression slopes assumption, and looking into the impact of data missing at random. This will be accomplished through a regiment of **Monte Carlo simulations** with 1 000 000 **iterations**, a choice discussed later.

One theoretical contribution I propose is the concept of **contingent power of the simple effect** when comparing a difference scores with an ANCOVA. A statistical method is only useful if it can answer the research question being asked. I argue that in mixed-design models, the **group** difference of an ANCOVA is semantically equivalent to a simple effect test contingent on the interaction being significant in a RMANOVA.

My initial hypothesis was that when the ANCOVA is capable of answering the research question, it is invariably equal or better than a difference score analyses in terms

of statistical power. This proved to be mostly true, but not universally so. The analytical model targeted is that of fully randomized mixed models, and excludes cases where the group assignment is pre-existing (e.g. gender). Non-random group assignment is excluded. ANCOVA should not be used with non-random group assignment cases because the pre-tests become dependent on group assignment. This means that we can no longer discriminate the source of the effect in the model. The thesis focuses on a perspective in which the grouping variable is the variable of interest in the observed change, not the simple passing of time. The simulations and most of the discussion is limited to a context with two time points (such as a pre-test and a post-test). I do discuss the implications of the two time point findings in the wider perspective of multi-time point studies later on.

Before I begin, there are a few notes that I wish to make to clarify aspects of the current work.

First, in light of the recent statement made by the *American Statistical Association* (Wasserstein & Lazar, 2016) on the old issue of the misuse and over-reliance on *p*-values, I feel that I must clarify my position. I am of the opinion that discussing the relative ability of a method to properly model data as a function of reduced Type II Error without sacrificing Type I Error stability does not conflict with the higher-level questions of when and how to use these tools as a whole. Power is a function of sample size and effect size. As such this study on power in small sample sizes is implicitly an examination of our ability to detect change as a function of effect size. These models, before being tests of

significance, are techniques to estimate parameters, and parameter estimation is the core of a New Kind of Statistics (Cumming, 2014).

Second, I would like to clarify why I use the term correlation when speaking of the relationship between Time 1 and Time 2, rather than unstandardized partial regression coefficient. In the particular circumstance where the measured and outcome values are on the same scale and the variance are equal, the unstandardized regression coefficient, the standardized regression coefficient, and the standardized correlation all become equivalent. This can be observed in equation 1, where $\hat{\beta}$ is the partial regression coefficient, *r* is the standardized correlation coefficient between the Time 1 and Time 2 data, and *SD* is the standard deviation of the respective time points.

$$\hat{\beta} = r_{T2_i,T1_i} * \frac{SD_{T2_i}}{SD_{T1_i}} \quad (1)$$

Although the final model used later also has a grouping variable, all data will be generated independently for both groups, meaning that the slope / regression coefficient used will still be a good approximation. Since the relationship between the two time points uses the correlation coefficient in the variance-covariance matrix, the term used will be that of correlation.

Lastly, note that there are multiple ways of writing the same model, and often multiple terms are used in the literature to refer to a given concept. For this reason, a glossary is provided on p.84 with the intent to clarify notational equivalence and synonymous terms. The definition and synonyms are placed next to the preferred term

used in this work. The first time a term is used, it is bolded to indicate that an entry is found in the glossary.

Chapter 1- Literature review

In this chapter, I review the important themes explored over the years. The review covers the topics by adopting a chronological structure. The reason for this choice is that the development of the ideas surrounding the use of baseline data as a control variable were discussed in a series of articles by multiple researchers over given periods of time, after which the debate matured and moved on to the next topic, or sometimes was left to sit as other aspects progressed, to be revisited later. This progression was heavily influenced by an increased access to computational power, and by the development of parallel fields which relied on this computational power (such as item response theory). By contextualizing the discussion within its time period, we get an insight as to why some of the discussions were concluded the way they were, and why, years later, researchers would suddenly go back to an older issue.

To understand this computational context, and to the credit to those who developed these ideas when computers were still in their infancy, we need to look no further than Donald Zimmerman's personal webpage http://mypage.direct.ca/z/zimmerma/ . He recounts running his first Monte Carlo simulation in Fortran on punch cards, and running it on a computer 1 200 times slower than his 3.4 ghz personal computer (no exact date is given for this post, but we can presume it is a single core Pentium 4; already 12 years old as of this writing); let alone the time required to code and re-code simulations in this fashion. By comparison, Intel now provides specialized kernels to optimize linear and matrix algebra on their multi-core CPUs. This means that the Intel Core i7 from 2014 that I am working from is

approximately 50 times faster than that Pentium 4; making it approximately 60 000 times faster than the computer Zimmerman used for his first Monte Carlo simulations.

1967: Lord's Paradox

In 1967, Lord published a two page editorial on the potential conflicts that can arise in using difference scores or a covariate on the same data (Lord, 1967); a situation that came to be known as Lord's Paradox. His editorial brought to the forefront a known issue, and ignited a debate on the appropriateness of ANCOVA versus difference scores. At this point, tensions on the topic had been rising for about a decade, and although this was not the first article on the shortfalls of either method (e.g. Lord, 1963), this editorial succinctly embodied the issue of choosing the right analytical method. By embodying the issue in such a short and accessible paradox, and as a leader in the field of test theory, Lord effectively kicked off the debate that ensued for the next 50 years.

In his example, Lord imagined two statisticians analyzing the same dataset. The situation consisted of two groups of school children (boys and girls) who would have their weight assessed at the beginning and end of the year. In this case, the means for the groups remain the same. The first statistician runs a difference score test, and finds that there was no change within the groups. He then concludes that there was no differential effect on the two groups. The second statistician runs an ANCOVA and finds that the boys gain more weight than the girls when controlling for the initial weight difference between the two groups. The second statistician concludes that there was a statistically

significant effect, concluding that boys gained significantly more weight than girls over the year.

1960s and 1970s: Difference Scores Criticized

During this period, the debate centred on the idea of problematic reliability of difference scores. Initially, this issue stemmed from concerns about the validity of difference scores, and was unrelated to that of the use of ANCOVA. The perceived issue with the difference scores was seized upon by some in the debate about whether to use a difference score or an ANCOVA, and the two issues were henceforth linked.

Reliability of a difference score is focused on the change aspect of a repeated measure. To give a definition, reliability of a score is the proportion of the variance of a score that is not due to error; the portion that persists when it is measured again. On the other hand, the reliability of the *difference* score is the application of this concept of reliability to the difference between the two source values rather than to the pre and post-test values themselves. For some, this controversy served as a *de facto* justification for the use of alternative methods, such as partial correlation (Wall & Payne, 1973). Sadly, this perceived weakness of the difference score is based on single study observations (Porter, 1962), overly simple simulation studies (Wall & Payne, 1973), and generalizations from overly constrained equations (Lord, 1963; Cronhach & Furby, 1970 as cited by Zimmerman & Williams, 1982).

Take for example, Wall and Payne (1973). This study was based on simulations for which a total of eight iterations were produced to compare two general cases (four per

case). They concluded that difference scores are deficient, and partial correlation should be used instead.

Meanwhile, Cronbach and Furby (1970) explore from a mathematical model perspective the use of difference scores and **residualized change scores** (methodologically a precursor to ANCOVA; no longer pertinent, but still indicative). They observed that the regressor method makes use of more available variance (better parameter estimates), but surprisingly concluded that researchers would be better advised to reframe their research questions and simply avoid gain scores.

Kutner, Nachtsheim, Neter, and Li (2005), who's book was originally published in 1974 under the first two authors, demonstrated that when the relationship between the Time 1 and Time 2 variable is perfect ($r = \pm 1$, and so $B_1 = 1$), the equation of an ANCOVA

$$\mathrm{T}_2 = B_0 + B_1\mathrm{T}_1 + B_2 G + \boldsymbol{\varepsilon} \quad (2)$$

is equivalent to the factorial ANOVA using a difference score (which they called a change-score model):

$$\mathrm{T}_2 - \mathrm{T}_1 = B_0 + B_2\mathrm{G} + \varepsilon \quad (3a)$$

where $\mathbf{T_1}$ is the Time 1 score, $\mathbf{T_2}$ is the Time 2 score, B_0 is the origin, B_1 and B_2 are standardized regression weights, **G** is the **grouping variable**, and $\boldsymbol{\varepsilon}$ is error.

This equivalence becomes obvious once you displace the $\mathbf{T}_1$ parameter to the other side of the equation and give the B_1 parameter a value of 1:

$$\mathrm{T}_2 = B_0 + \mathrm{T}_1 + B_2\mathrm{G} + \varepsilon \quad (3b)$$

Given that the factorial ANOVA using a difference score produces identical results as a 2 × 2 mixed RMANOVA, the equivalence between the difference scores analysis and the ANCOVA can be extended to a RMANOVA. They argued that when the regression coefficient diverges from unity, the ANCOVA can be better.

Overall and Woodward posited in 1975 that pre and post-test score reliability is more important than the difference score reliability. They point out that the observed null reliability of the difference score, that is, of the combination of the pre and post-test scores, was a quirk of no consequence. They conclude that since the reduction in reliability of the difference score is related to an increase in the correlation between the time points, the difference score's validity is unimportant when compared to the increase in statistical power that accompanied this stronger correlation between Times 1 and 2.

1980s and early 90s: Incremental Theoretical Work

During the 80s and early 90s, the debate advanced incrementally with much back and forth. By the end of this period, it was clear that the difference scores could not be dismissed based on the supposed reliability issue. It also became clear that better modelling of error variance was needed to settle the reliability issue with difference scores.

Johns (1981), like Overall and Woodward (1975) stated that, when looking at difference scores, reliability of the difference score is rarely of interest. He stated that the lack of reliability of the difference score is likely to be problematic only in multivariate scenarios where the difference score is analyzed in tandem with other variables that are

related to the difference score or its components. He then listed a series of examples where difference scores were obtained from a difference between various non-repeated measures to obtain new outcomes, or using difference scores as well as their components in a same model. These concerns can be summarized as concerns about important methodological violations that are unrelated to the inherent qualities of the misused methods.

In 1982, Zimmerman and Williams contributed the first of many articles on the reliability of difference scores on which they, and eventually Zumbo, will work. They underlined two core issues with the equation used to calculate difference score reliability.

The first issue that Zimmerman and Williams (1982) highlighted is the assumption that the error variance of the components is null, and as such, that the relationship between these errors ($\rho_{E_{T2}E_{T2}}$) is also null. This means that a reduced equation which does not account for the relationship between the Time 1 and Time 2 errors was used in prior articles outlining the issue of difference score reliability. The second issue is that the variance of the Time 1score $\sigma_{(T1)}$ and variance of Time 2 score $\sigma_{(T2)}$ are often assumed to be the same. From this erroneous assumption follows the conclusion that the ratio of the two (λ) is of 1, when that is not the case; dropping yet another term from the reliability equation in which λ is a multiplier of the correlation of the Time 1 and Time 2 reliability coefficient ($\rho_{T1T1'}$, $\rho_{T2T2'}$).

By using the standard equation (4), in which ρ_{T1T2} is the correlation between the Time 1 and Time 2 scores

$$\rho_{\Delta\Delta'} = \frac{\rho_{T1T1'} + \rho_{T2T2'} - 2\rho_{T1T2}}{2(1 - \rho_{T2T2'})} \quad (4)$$

and allowing λ to vary, the equation (5) becomes more complete, in which we can see where λ is dropped when it is assumed to equal one.

$$\rho_{\Delta\Delta'} = \frac{\lambda\rho_{T1T1'} + \lambda^{-1}\rho_{T2T2'} - 2\rho_{T1T2}}{\lambda + \lambda^{-1} - 2\rho_{T2T2'}} \quad (5)$$

Finally, if $\rho_{E_{T_1}E_{T_2}}$ (the correlation between the error terms of the Time 1 and Time 2 scores) does not equal 0 (equation 6), we can see how the simplified equation came to be.

$$\rho_{\Delta\Delta'} = \frac{\lambda\rho_{T1T1'} + \lambda^{-1}\rho_{T2T2'} - 2\rho_{T1T2} + 2\rho_{E_{T_1}E_{T_2}}[(1 - \rho_{T1T1'})(1 - \rho_{T2T2'})]^{1/2}}{\lambda + \lambda^{-1} - 2\rho_{T1T2}} \quad (6)$$

With this expanded equation, reliability of the difference score was no longer leaning towards a null value. Furthermore, they pointed out that adding highly reliable gain scores to pre-test scores can change the variance of the post-test scores which, in turn, could then actually serve to enhance the reliability of difference. In essence thus means that in a **fan spread growth model**, the difference scores' reliability can be higher than that of the initial scores. Ultimately, the conclusion is not that difference scores are usually reliable, but simply that they can be. As is always the case with statistical analyses, using appropriate methods and assumptions is key.

Allison (1990) defended the use of difference scores as dependent variables, arguing, rightfully so, that the debate had centred on the rejection of difference scores for their unreliability. According to Allison (1990), the two major issues that researchers

seem to have had with change scores involve their reliability and the potential for regression to the mean from pre- to post-test.

Allison (1990) explored variations on an example centred on the situation described in Lords' Paradox (1967). Using data comprised of 48 participants with pre-existing group assignment (18 in the treatment group, 30 in the control group), he looked at the impact of treatment on two variables. For the first variable, where there was seemingly no change, Allison illustrated that the group variable had a statistically significant relationship with Time 2 scores. For the second variable, where there was an interaction, the regression model failed to underline a treatment effect, while the difference score method yielded a statistically significant relationship between groups and difference scores. These contradictory results were caused by the dependence of pre-test score on the group assignment. Consequently, the regression approach under-adjusted the scores. Allison went on to demonstrate this point by comparing two data-generation models, one consistent with the regression approach and one consistent with the change score approach. When data generated by the latter model are analyzed using the regression approach, bias is necessarily introduced in the estimation of the treatment effect whenever initial pre-test differences exist between the treatment and control groups (thus resulting in Lord's paradox).

Allison (1990) also listed a series of hypothetical situations to highlight circumstances under which one method would be more appropriate than the other. He concluded that when treatment assignment is based on the pre-test score, or when Time 1 has a "true causal effect" on Time 2, ANCOVA is superior, with the caveat that the

difference score method is superior when group assignment is based on pre-test but that there is no treatment effect. This last part of his conclusion seems somewhat confusing since it assumes prior knowledge of the end results.

Wainer (1991) gave an example in which animal heart rate was measured, and concluded that because we can assume that the heart rate would have remained unchanged without the intervention, we can attribute the change entirely to the treatment. He argues that in such situations, we can use a difference score analysis. His general conclusion is that in other cases, we cannot tell which method is most appropriate as either can be biased.

Zimmerman et al. (1993) demonstrated that, contrary to what is said in other articles criticizing the use of difference scores based on their reliability, reliability is not always directly related to power. It implies that, even if we ignore the corrections to difference score reliability that Williams, Zimmerman, Zumbo, and their collaborators have proposed, low reliability in difference scores does not necessarily imply that this test will have low statistical power. This conclusion further dismisses the idea that the difference score's apparent reliability problem was a reason to suspend all use of the methods relying on them.

Williams and Zimmerman (1996) presented a series of modifications to the way in which validity is adjusted in an attempt to account for the importance of the correlation between the errors of the difference scores' components. The conclusion remains that difference scores can't be ruled out simply because of the statistical argument on

reliability, as they demonstrate that this argument can go either way depending on the nature of the relationship between components of the difference score.

Mid-90s to Early 2000s : Comparison of the Methods

The mid-90s to early 2000s set the stage for a series of simulation studies. William, Zimmerman, and Zumbo are now the leading voice of caution against unequivocally disavowing difference scores, as they keep refining their work on the reliability of difference scores. Meanwhile, except for Overall and Doyle's (1994a) avant-garde simulation work, most of the debates are still occurring at a theoretical level. One element that could explain why there was such a difference in Overall and Doyle's approach in contrast to their contemporaries is that the former are from an applied clinical setting, while the latter are mostly from a psychometric background. During this period, researchers are putting more weight on comparing the difference score and ANCOVA rather than studying them in isolation.

Overall and Doyle (1994a) ran an ambitious set of Monte Carlo simulations on the topic of ANCOVA and difference score's power, with 10 000 iterations per parameter adjustment. The goal of the study was to establish the sample size required to reach a certain power, and to compare the results with what the power formulas predicted. They used four, six, or eight time points in the repeated measure variable, and used the relationship between the end-point and baseline as the parameter that they varied for r = .3, .5, and .7, with a mean change of *Cohen's d* = 0.5. Although they did vary the

mean difference as well, it was only in the formula approach, not in the simulations. They used an exponential decay correlation matrix for their data.

Overall & Doyle (1994a) looked at both the overall interaction of the multi-time point data, which they called endpoint score, as well as the linear trend component of the overall interaction. They argued that linearity would indicate the absence of an interaction, as the linear trend accounts for most of the variance in the interaction in a repeated measure model. In this situation, the data was also simulated in a way in which it does not plateau, making the use of the linear trend score acceptable as a generalizable substitute. Furthermore, the linear trend would be uniquely useful to detect a plateau effect when there is one, a point that is especially important given that their background was in medical research.

Overall and Doyle (1994b) advocated for the use of baseline data as a covariate as a corrective measure to linear trend direction (effect across time) in randomized mixed designs, and for within-subject effects by using composite trend scores instead along with the covariate. They highlighted the corrective impact of a pre-test covariate on some of the directional biases that can be observed when baseline scores are congruent or incongruent with change.

Maris (1998) suggested that when there are missing data at post-test, if assignment is based on the pre-test, the ANCOVA method is preferable, but that if the assignment mechanism is unknown, we cannot know which estimator will be biased. He also goes to some lengths to clarify the issue of the regression towards the mean phenomenon. Some researchers had assumed that it had some relevance to the principles

behind using an inferential regression, but as Maris points out, it is simply an observable phenomenon that occurs as a result of sampling, and is not dependent or related to the research question at hand. It can occur in situations that are associated with a biased difference score or not. Lastly, he states that measurement error cannot be a reason for bias in the difference score estimator if the assignment was not affected by the pre-test.

Owen and Froman (1998) focused on what they perceived to be illegitimate uses of ANCOVA. They focus on two cases for which they criticize the use of ANCOVA. The first is the practice of substituting variables that may be related with the other independent variables as a proxy in an analysis; stating that this leads to potential issues of bias and loss of validity. The second case is the use of covariates that serve a proxy role (i.e. assuming that one variable can be used as a covariate under the presumption that it is sufficiently similar or related to the factor that the researchers intend to control) and are not sufficiently correlated with the dependent variable. They go on to list extra precautions that must be taken. The first is to minimize measurement error in the covariate as its presence reduces power, and can increase Type I error in rare instances. The second is to avoid severe violations of homogeneity of regression slopes (unequal $r_{(T1,T2)}$ between groups). And lastly, one must ensure that the relationship between the covariate and the dependent variable is linear.

Zumbo (1999) provided a thorough overview of the issue of measuring change up to date. He also summarized the reason for the issue of measuring change in two sentences.

> *But in behavioral research, unlike the life and physical sciences, we seldom know in advance just what factors govern the phenomenon under study, nor how to accurately measure even those that we do suspect. Therefore, the measurement and analysis of change has been a persistent problem for social/behavioral science statisticians, psychometricians, and empirical researchers for more than 60 years.* (page 270)

One aspect that Zumbo brought to light in this debate on the reliability of difference scores is that when the correlation between the time points are negative, reliability of the difference score is higher. He reiterates that when Time 1 and Time 2 variances are equal, difference score should not be used (as allowing the two to vary is the key in the cases demonstrated by Williams, Zimmerman, and Zumbo where the reliability of difference scores was adequate).

Cribbie and Jamieson (2000) ventured into the resolution of Lord's Paradox. They stated that with a regression coefficient of less than one between time points, "the observed post-test difference will be greater than expected"; which they attributed to the regression towards the mean effect, an issue discounted by Maris (1998). They also stated that ANCOVA should only be used with randomized group assignments and not with pre-existing groups.

Cribbie and Jamieson (2000) criticized the work of Maris (1998), and Wainer (1991), which is based on Rubin's 1988 model (as cited by Cribbie and Jamieson, 2000), saying that it detract from the real issue. Cribbie and Jamieson (2000) criticized the focus on establishing what would have happened had the groups not been treated (i.e. been the

control group instead), rather than determining the reason for the baseline differences. Referring to their earlier work, they reiterate that “ANCOVA is biased to find significance in means which stay apart (parallel lines, as in Lord’s paradox) or which diverge” (page 897). They call this the regression bias. They go on to state that when a variable is uncorrelated with the pre-test score, both tests yielded comparable values. However, when a variable is correlated with the pre-test score, the ANCOVA is biased to detect a statistically significant relationship with change. This bias takes the form of an increase in an increase in rejections of the null hypothesis, including Type I errors, when the change is in the same direction as the correlation between time points (+/+ or -/-), and an increase in Type II error rates when the direction of change and the correlation between time points are in opposite directions (+/- or -/+). Lastly, they state that when the correlation between a “third variable” (a continuous variable that replaces the group variable) and change is 0, large correlations with baseline resulted in large Type I error rates. They added to this “in every case where there was a large correlation of the third variable with baseline, regression was biased to detect correlations with change in only one direction” (page 898). Ultimately, their conclusion was that this is because of error that exists in the pre-test measurement. They showed through simulation that structural equation modelling (SEM) should be used as it is immune to the outlined issues, because error terms are explicitly modelled in SEM.

Oakes and Feldman's 2001 article is an exploration of the difference score versus ANCOVA models in which they calculated statistical power. Their main conclusion is that in randomized experiments, ANCOVA is unbiased and has more statistical power

than the RMANOVA. This finding is followed by a somewhat contradictory statement in which they say that the reason for the extra power is the "untenable assumption that pre-tests are measured without error" (page 18), and that when there is measurement error the difference scores may be equally or even more powerful. They do warn, like others before them, that in quasi-experimental designs, in the presence of measurement error, "parallel regression lines become flatter and may decrease or increase their vertical separation, depending on the positions of the treatment and control data" (page 9). They find that ANCOVA is about 30% more precise in estimating parameters when perfect pre-test reliability is assumed in a randomized experiment, but has about 30% less precision in quasi-experimental situations, whereas the difference score would show little or no bias.

Miller and Chapman (2001) take issue with what they consider to be the "widely misused approach to dealing with substantive group differences on potential covariates" (page 40). The critique is an attempt to summarize in a non-technical fashion the issues that can arise from the misuse of ANCOVA, and a definition of those misuse cases. The idea centres on using covariates that remove too much explained variance in quasi-experimental designs, or that violate the homogeneity of covariance. In other words, the covariate cannot control for group differences on the covariate itself. For this, they quote Cochran (1957) in saying "it is important to verify that the treatments have had no effect on the covariate…. [otherwise] a covariance adjustment may remove most of the real treatment effect" (page 42). However, they do go on to cite exceptions specifically for the

assignment to group based on pre-test score, but consider this to be an ambiguous area that requires careful attention and interpretation.

Edwards (2001) examines what he perceives as being 10 misconceptions about difference scores. His suggestion is to replace difference scores with "polynomial regressions which use components of difference scores supplemented by higher-order terms to represent relationships of interest" (page 265). To do this he tries to objectively dispel myths in favour and against the use of difference scores, as any fundamental misunderstanding can lead to the misuse of a statistical method. Below are the ones that are relevant to this thesis. They are listed with the italicized myth, followed by a summary of his response.

1. *Low reliability is the difference scores' only issue.* For years, there was a disproportionate focus on the lack of reliability of difference scores. With the reliability issue having been shown to be overblown, some used this as a justification for the use of difference scores, rather than comparing them with alternative methods.
2. *Difference scores provide a conservative statistical test*. This is a fundamental misunderstanding of statistical analysis. An inflation of Type II errors is not a "good" form of conservatism in a test.
3. *Measures that elicit direct comparisons avoid problems with difference scores*. This merely shifts the onus of creating the difference from the researcher to the respondent, is a double-barrel question, and incurs biases by

demanding further cognitive processes from the participants, on which individuals may differ.

4. *Categorized comparisons avoid problems with difference scores*. This merely accentuates the loss of information by categorizing continuous measures and inherits the other issues associated to difference scores.
5. *[Use] product terms as a substitute to difference scores applied in a hierarchical multiple regression analysis*. However, this makes it impossible to assess the curvature in the relationship and so cannot assess the congruence effects.
6. *Hierarchical analysis provides a conservative test of difference scores*. This method alters the relationship it is intended to capture, and can distort the outcome.
7. *Use polynomial regression as an exploratory, empirically-driven procedure*. As with any other method, the exploratory or confirmatory nature of an analysis is in its use, not its nature. Polynomial regression can answer the same questions as difference scores, except that *a priori* hypotheses of the latter can be explicitly tested in the model.

Edwards concluded that ultimately, since most difference scores are a special case of polynomial regression, the debate of using a polynomial regression versus a difference score is moot. Polynomial regression does have some limitations; complex models can contain many terms, requiring larger samples, and inherit the multiple regression models' assumption of being collected with no measurement error.

Mid-2000s to Present : Simulations

Ultimately, William, Zimmerman, and Zumbo did call for the need for simulation work. As is usually the case in a good debate, both sides did have valid points. However, perpetually illustrating potential issues and strengths of both methods ultimately lacks the relative weight of these strengths and weaknesses. Although the background discussion remains relevant and is needed to move forward; manipulating various parameters of concern to observe the relative importance they hold is the only way to know which concerns are more important than the next when so many of those concerns have already been shown to be valid in a given set of circumstances.

Van Breukelen (2006) took a differential analysis approach, with the goal of exploring the purpose and limitation of the two methods. The work was largely theoretical, but was supported by an example using an existing clinical dataset, which was resampled to meet certain conditions. From a total of 180 participants (148 with no missing data), he resampled the data down to groups of 20 participants to test the effect of pre-test group allocation versus pre-existing group allocation, after biasing the experimental group by increasing their scores. This method is very limited in scope and validity, but it is one of the few studies that tried to wade into the topic of pre-existing groups, let alone bring in any type of manipulations, as limited as they may have been. From this, he created a situation where, based on sampling and the mean modification, there was a difference between the tests. He claimed that with pre-test score based assignment, the difference score method produced a Type I error while ANCOVA did not, and that the opposite was true with the pre-existing groups. He closed his article by

positing that the larger the pre-existing group difference is, the worse the results will be for the ANCOVA.

Wright (2006) focused on Monte Carlo simulations. Each simulation had 1 000 iterations per parameter adjustment; sample size was always of 100; and effect size as described by $\boldsymbol{\beta}$ (the unstandardized regression weight) was varied with values of 0, .5, 1, and 2. Wright deemed that an estimate was unbiased if it was within two standard errors of the expected value.

In Wright's first simulation, group assignment was determined by pre-test scores. He uses the inter-quartile range (IQR) as an indicator of power due to the high Type I errors when there is a small bias, where a smaller IRQ is supposed to be associated with more power. Wright found that the t-test had smaller IQRs, but that the t-test's bias increased as a function of measurement error.

In Wright's second simulation, baseline scores are associated with, but do not determine group allocation (he calls this group *allocation based on ability*). The group allocation is accomplished by having a third variable that is related to the baseline scores and the grouping variable. That is, the group assignment is accomplished by probabilistically assigning cases to groups based on ability, but not as a direct result of pre-test scores. From my understanding, this type of allocation is similar to a pre-existing group difference, but again, group allocation is not exclusive to this third variable; so it is not quite a pre-existing group allocation. Again, as measurement error increases, the IQR increases, but the t-test is unbiased while the ANCOVA is biased. This bias increase as

the measurement error increases, but is not sufficient to cause Type I error problems according to Wright.

The second simulation is replicated with several variations. In the first replication of the second experiment, slopes are allowed to vary (not parallel). In this case, the t-test has the larger IQR, but this only becomes apparent with the larger effect sizes. In the second replication of the second simulations, random assignment is introduced in the simulations. In this case, as measurement error increases, both IQRs increase, but the ANCOVA is more powerful and both tests remain unbiased.

In a third replication of the second simulations, Wright incorporates elements of the first simulations. People are assigned to a condition based on "ability" as with the second simulation, but this is based in part on test scores that have the same measurement bias as the pre-test. This makes the assignment somewhat of a hybrid of the first and second simulation, and is run with 10 000 iterations. In this case, both methods fail to properly assess true differences. They both produce a group effect where there was meant to be none.

For the third simulation, effect size is allowed to vary with ability so that if random allocation was used, there would be an interaction between group and pre-score. In the first set, group allocation is based on the pre-score, and the t-test is biased for the group effect and the interaction while the ANCOVA is not. For the first replication, group allocation is random, and all estimates are biased for group effect with fixed or varied slopes. For the third and last repetition, group allocation is based on ability. When the slopes are the same, the t-test estimates are unbiased for the group effect. When the

slopes are allowed to vary, the t-test produces biased results. The ANCOVA on the other hand produces biased estimates in all cases. Wright's closing conclusion is that when possible, randomized group allocation should be used.

Zimmerman, in 2009, published a fairly expansive and extensive article that focused on the idea of simulating a population, resampling from it, and then evaluating the reliability of the components and of the difference scores, knowing the population parameters from which the samples came. In unison with his past conclusions, he reiterates that difference scores can sometimes be reliable. However, one element that does stand out is that he tested this using various modified reliability equations, each variation reflecting a different assumption on the variability of the scores. From this, he concluded that each method can yield "quite different values" (page 41). Again, this work demonstrates that researchers can justify using difference scores if they have thoroughly justified the variance model used. I would like to point out that this thorough justification is probably beyond the interest and ability of most researchers (although maybe not as much in the field of testing and measurement from which Zimmerman speaks).

Petscher and Schatschneider (2011) ran sets of Monte Carlo simulations, with 1 000 repetitions per parameter variation to compare ANCOVA with a difference score model in a randomized experimental design. They focused on changing sample size, normality of the pre and post-test distributions, the correlation between pre-test and post-test, and correlation between pre-test and post-test change. Their main contribution was to discuss the way in which variance changes, as expressed by the correlation between the pre-test score and the difference score, affects the outcome of the statistical tests. A

negative correlation between the pre-test score and the difference score is known as *mastery learning*, that is the improvement was strongest with lower scoring individuals; a null correlation is known as ***parallel growth***; and a positive correlation is known as ***fan spread growth***. The correlation values used to model each of these three growth types were -.3, 0, and .3 respectively. The data generation model they chose constrained their ability to vary the Time 1 Time 2 correlation freely, and so values of .2, .6, and .8 were used for fan spread growth, .4, .6, and .8 for parallel growth, and .6, and .8 for **mastery growth**. Sample size was also varied from 40, 60, 100, 400, and 1 000 participants. Skewness was varied between .5, -.5, and 1.

Petscher and Schatschneider stated that both tests respected the .05 Type I error rate, but failed to provide further details. They go on to state that ANCOVA generally has more power, but that the difference shrinks as *N* increases and as the Time 1-Time 2 correlation increases, that skewness has very little effect on power, and that in fan spread growth models, ANCOVA loses its relative advantage. Ultimately they conclude that the effect comes down to the fan growth / mastery growth parameter manipulation more than anything else.

Lastly, Petscher and Schatschneider also performed a review limited to two journals to evaluate the method of analysis used in growth studies from 2002 to 2007, and found that of the 61 studies found, 27 use a randomized pre-test-post-test experimental design, and of those 12 used ANCOVA, 10 used RMANOVA, and 5 simply compared the post-tests. In none of the articles selected was there a justification for the choice of the statistical analysis performed.

Thomas and Zumbo, in 2011, published another article on the issue of the reliability of difference scores, expanding their prior work to RMANOVA (in this case, substituting the variance component with the MSE), attaching the reliability to the non-centrality parameter of repeated measure test. They once again demonstrate that when variance is properly accounted for in the reliability calculations, the difference score can be reliable.

Kisbu-Sakarya, MacKinnon, and Aiken's 2012 article is the latest using Monte Carlo simulations to explore the differences between ANCOVA and difference scores at this time. They also included residual change scores, but these are of little interest given that residual change scores are not valid, as they are equal to ANCOVA estimates in the best of conditions, but these estimates tend to be closer to 0 compared to the ANCOVA's estimates, as baseline imbalance grows and sample size diminishes (Forbes & Carlin, 2005). They varied the pre-post correlation (.3, .5, .7 and 1), stability as defined by β (0, ±.14, ±.39, ±.59), as well as reliability (ρxx, ρyy = .5, .8, and 1) sample size (100, 200, 400, and 1 000), as well as a small and large baseline imbalance. Baseline imbalance is manipulated ±.14 or ±.59 to the pre-test scores that were in the treatment group.

Kisbu-Sakarya et al. (2012) reported Type I error rates as being approximately .05, and similar between methods with nominal values between .04 and .07. As for power, they concluded that under baseline balance conditions, ANCOVA was always superior when stability was not perfect, but that results were equal when stability was perfect. Under baseline imbalance condition, they found that ANCOVA produced higher Type I error rates when reliability was not perfect, and that power was only higher for the

ANCOVA when the baseline imbalance was positively related with the overall score change (either treatment had a higher score and change was positive, or treatment had a lower score and change was negative). The results are interesting, but with only 1 000 iterations per variation, an insufficient number of iterations were run. The power results are discussed in an overall qualitative fashion; this probably stems from the variability in the results and the very large number of tables. Moreover, in these tables, power seems to vary as much if not more within a test than between tests (trends are not stable). One of the limits that they themselves bring up is that the data were strictly modeled from an ANCOVA model, which could be self-confirming.

Review of the simulation studies

In this section, I concentrate on the articles where simulations were performed. This is essential to highlight what still needs to be explored. Table 1 synthesizes the simulations that were found on the topic of the relative power of difference scores and ANCOVA. The term simulation is taken loosely given that in one case only one repetition of each scenario was conducted.

There are three main column headers under which are nested sub-columns. The first indicates the group allocation method. The second is a list of the parameters that are varied within the simulations. Blank boxes simply mean that the values are not reported and are almost certainly left to vary randomly or were measured as an outcome rather than manipulated at the outset. When explicit values are not simple to include (e.g., Kisbu-Sakarya et al.'s 2012 measure of baseline imbalance), further details are provided

in Chapter 1 when discussing the author's paper. The last group of columns is a summary of the results found. The goal is to present as best as I can what researchers found in terms of Type I and Type II error rates.

From this table, we can see that the following elements are missing from the analyses:

- Systematic manipulations of parameters. There seems to be a fairly substantial lack of studies when the parameters are at extremes; simulations are often limited to one extreme or the other, but not both.
- The numbers of repetitions for many levels of the parameters are too small. Although some specific parameter settings were tested at the minimally acceptable 10 000 iterations. This leads to too much variance in some estimated values. Additionally, error ranges (e. g., standard errors, SE) are entirely lacking, with conclusions sometimes being drawn questionably form these differences.
- Type I error: The compound effect of too few iterations and no SE makes it difficult to properly interpret certain data such as the Type I error rates; the evaluation of Type I error rates is also generally lacking in thoroughness, which leads to a questionable discussion of bias.
- Missingness is completely ignored in the simulations, although discussed in the literature as a potential theoretical issue (Maris, 1998).
- Violation of the homogeneity of regression slopes principle is sometimes mentioned as a concern, but has not been thoroughly explored.

Chapter 2- The Contingent Probability Proposition

Theory

When a statistical test is chosen, before even looking into the issues of assumptions and power, the most basic question is "can this analytical method provide relevant insight into my research question". And so, in exploring the issue of relative power of the statistical methods, we must not lose sight of that fundamental question. "What information does ANCOVA provide when it is applied to a mixed design study?"

Interpreting the results of an ANCOVA, when a group difference is statistically significant, we can infer that there was a treatment effect. The change was differential between the two groups, even if both groups may have changed. This information can be equated to the type of information that the interaction term of a traditional difference score analysis yields.

What I propose is that although the RMANOVA interaction (or factorial group difference on difference score) is mathematically equivalent to the ANCOVA group difference under conditions of a perfect correlation between Times 1 and 2, we can infer further information from the ANCOVA's group difference. What does a differential effect mean when we are dealing with **estimated marginal means** (EMM)? The statistically significant difference of the group EMM gives us the difference between both groups at the second time point, assuming that the groups were equated at the covaried time point. In the case of a two time point model, this means that we are also getting a response to the question which the statistically significant interaction term of the mixed-

model analysis would beg, that is, “what are the simple effects, if there are any?” This is because the ANCOVA model frames the question in a context of groups that were statistically equated at Time 1. Therefore, with an ANCOVA, we get the answer to the question “is there a treatment effect?”, but through the EMMs, we also get the answer to the question, “what is the Time 2 difference in a hypothetical context where we know that there is no statistically significant difference at Time 1?”

Implications and Application

Type II Error and Power

In the studies that follow, I am always comparing ANCOVA to a RMANOVA so that there is a benchmark for comparison. When testing for power, I have chosen to compare the group difference of the ANCOVA with the contingent power of the RMANOVA’s simple effects. What I mean when I say contingent power of the simple effect is that I use the effective power of either simple effects, regardless of whether it is at Time 1 or Time 2, contingent on the interaction term having first been found to be statistically significant. The simple effect is the difference between the group means, at Time 1 or Time 2, using the interaction error term as the denominator. The reason for this choice is that a researcher who is met with a non-significant interaction term should not move forward, and had there been a true simple effect, they would then have fallen prey to a Type II error. It is clear, however, that the interaction term provides the direct equivalence of comparing whether or not there was a differential change. Therefore, the

interaction term's significance level is also collected and analyzed, and will be discussed to assuage concerns regarding the validity of these comparisons.

It is also important to note that the pooled error rate for the simple effects was also explored, but as it creates a stricter test (the error term becomes inflated, reducing the number of rejections of the null hypothesis), and it is not easily accessible, it was not selected as the error term of choice. There was both a concern of biasing the analyses against the RMANOVA, as well as the fact that most researchers would not program in their own error terms to perform their analyses. Therefore, regardless of the recommendations made by some in this regard (Winer, Brown, & Michels, 1991), the pooled error term will not be used.

Type I Error Rate

In the simulations, Type I error rate is measured by tallying rejections of the null hypothesis when the population is known not to differ. When I measure the Type I error rate, I do not use the contingent probability; I use the interaction term's test of significance. I also limit myself to the interaction term difference, and ignore the other Type I errors that a researcher may encounter. When testing the no difference condition, I did find that the Type I errors compounded with the three tests; group difference, time difference, and interaction of the group and time. Any of the main effects can produce a Type I, and any of them would mislead the researcher. The contingent test would be stricter and reduce the Type I error rate, but since other possible Type I errors are being ignored, the combination of the two elements would make a meaningful and practical interpretation difficult.

Limits

There are some limits to the use of a contingent power measurement. Its relative power advantage is diminished when more than two time points are of interest, as there would still be an interaction term with $n - 1$ time points. Furthermore, if large higher order interaction terms are decomposed, we have to part out the α between the tests.

Chapter 3- List of Experiments

Generic Methodology

To explore the various questions related to the difference between RMANOVA and ANCOVA, I chose to use Monte Carlo simulations. This approach consists of repeatedly creating random samples, and analyzing these samples. Although each experiment has its own particularities, there is a general underlying method upon which each subsequent experiment builds.

Scenarios were run in parallel by using 20 server instances (DigitalOcean, 2016), each with the processing power equivalent to a single core of an Intel Core i5 3.2 GhZ processor. Using the free RStudio Server license (RStudio, 2016) allows for a single instance or R (R Core Team, 2016) per computer, hence the limit to a single core per server instance. Using a pre-configured snapshot, server instances were opened simultaneously, and deleted once the scenario was completed, keeping the cost to a minimum (0.14$ USD/h total for all 20 instances). The DigitalOcean team did provide enough credits for all simulations to be run at no cost to myself.

R is used to generate and analyze datasets in a 2 × (2) design with a known mean, sample size, standard deviation, and correlation between Time 1 and Time 2. The code for the main sets of experiments is available in Appendix 2 and the code for experiment 2 is available in Appendix 3. The bivariate distributions are built using a method that reflects the RMANOVA model. This is in contrast to a method where *n* true scores would have been generated, and a variance model applied to each score. I chose this as I felt that

it allowed the data and sources of variance to be more objectively independent as I programed it. However, this also means that the variance becomes pooled in the *r* coefficient, and so does not distinguish between sampling error, measurement error, or any other source of error.

For each simulated participant in each group, a pair of Time 1 and Time 2 (i.e., pre and post) observations were drawn from a bivariate normal distribution with Time 1 and Time 2 means of 100, Time 1 and Time 2 variances of 15^2, and a Time 1 and Time 2 covariance of $r_{(t1,t2)}$ 15^2 (with $r_{(t1,t2)}$ being the correlation between the Time 1 and Time 2 observations). Treatment effects (*d*) were then modeled by raising the Time 2 mean in Group 2's bivariate distribution as indicated in the following experiment descriptions. The normal condition is to have $n = 20$ participants per group.If we look at equation 2 from Chapter 1,

$$\mathrm{T}_2 = B_0 + B_1\mathrm{T}_1 + B_2G + \varepsilon \qquad (2)$$

the parameter $r_{(t1,t2)}$ and the difference between Time 1 and Time 2 scores dictate B_1, and the difference between the groups dictates B_2. As the methodology for each experiment is provided, the parameters targeted by the experiment are bolded. Note that I do not intend to imply with the use of equation 2 that the data is generated using this model. I use equation 2 to illustrate which part of equation would be most affected by the data as it is generated for a given experiment, were it to be fitted to the regression model.

Each scenario is iterated 1 000 000 times, and each iteration is performed with a new randomly generated dataset, with the seed left to float. The result of significance testing is computed as a pass/fail with a threshold of $\alpha = 5\%$ and represented as a

proportion of successes/failures (depending on the context). This choice of 1 000 000 iterations is based on an article written in late 2015 to explore the type I error rate and statistical power as a function of mean change and strength of association between time points. With the initial 25 000 iterations, I found that sampling error was still sufficiently important to make the interpretation of data less evident; and as such less accessible to readers. With 10^6 replications per point, I chose not to report standard errors going forward.

To assess the two methods, I explored the impact of a series of parameters on power by manipulating them in pairs. The parameters explored are the correlation between Time 1 and Time 2 data, sample size, effect size, violation of the homogeneity of regression slopes assumption, large systematic baseline differences, and randomly missing data. Although I am repeating many scenarios that were independently assessed by other researchers in the past, I will not be looking at the impact of non-normality, as I feel that Petscher & Schatschneider's (2011) assessment that there was no impact in the data tied to the changes in skewness or kurtosis was sufficiently compelling to justify no further development. Not only does it seem pointless given the prior demonstration, but the addition of these parameters would increase the number of scenarios by at least 9 folds as all main permutations would have to be observed.

A caveat to keep in mind; whenever I specify a correlation value of 0 or 1, it in fact refers to a value of 0.001 and 0.999. When the covariance matrix is set to have a perfect correlation of $r = 1$ or $r = 0$, it causes issues and the models collapse (divisions by 0, complete absence of variance, etc.).

In the following subsections, I give specific details regarding each experiment performed.

Experiment 1: Varying the Effect Size Between Times

The goal of the first experiment is to observe the impact of a varying effect size (Cohen's *d* is used) for a full range of values of $r_{(t1,t2)}$ on null hypothesis rejections in the forms of power and Type I error. I hypothesize that there will be no important differences in Type I error, and that the ANCOVA will be a more powerful test than the difference-score test. I further hypothesize that supplementary power of ANCOVA will be inversely related to the strength of $r_{(t1,t2)}$.

The first experiment assumes random participant allocation; $r_{(t1,t2)}$ is varied from 0 to 1 by increments of .1; *n* is fixed to 40 (20 per group), and later to 200 (100 per group); effect size is varied by increasing the group 2 Time 2 mean from $d = 0$ to $d = 1.2$ by increments of 0.1 (i.e., unstandardized changes of the mean in steps of 1.5).

In the general equation 2 presented earlier, the manipulation of the effect size represents the difference between Time 1 and Time 2 in terms of standard deviations. The fact that this change will only occur in one group will be reflected in the B_2 term. The correlation between the Time 1 and Time 2 data, $r_{(t1,t2)}$, is expressed by the B_1 term.

$$\mathrm{T}_2 = B_0 + \boldsymbol{B_1}\mathrm{T}_1 + \boldsymbol{B_2}G + \varepsilon \qquad (2)$$

Experiment 2: Determining the Sample Size Required to Obtain .80 Power

The goal of the second experiment is to see what sample size is required in order to reach a power of .80 for each method, with $r_{(t1,t2)}$ restricted to a value of 0.5.

I hypothesize that the ANCOVA will require a smaller sample size as a direct consequence of its higher power.

The second experiment assumes random participant allocation; $r_{(t1,t2)}$ is fixed to 0.5; *n* per group is varied from 10 to 200 by increments of 10 (total sample of 20 to 400); effect size is varied $d = .1$ by increasing the group 2 Time 2 mean by 1.5, from 0 to $d = .9$. This range of *d* values is based on prior knowledge from the earlier experiment.

Experiment 3: Violating the Homogeneity of the Regression Slopes Assumption

The goal of the third experiment is to test whether the homogeneity of the regression slopes assumption has an impact on the power of either method. This is accomplished by having a $r_{(t1,t2)}$ coefficient that varies by group.

I hypothesize that the ANCOVA method will retain its superiority even when the homogeneity of the regression slopes assumption is violated, contrary to Owen and Froman's (1998) claim.

In terms of equation 2, varying the slopes means that there would be an interaction between the grouping variable (G) and the covariate (T_1). Hence, correlation between the Time 1 and Time 2 data, $r_{(t1,t2)}$, which is expressed by the B_1 term, will differ by group. However, this is not modeled in any of the equations used, which is why it is a potential source of difficulties of interest.

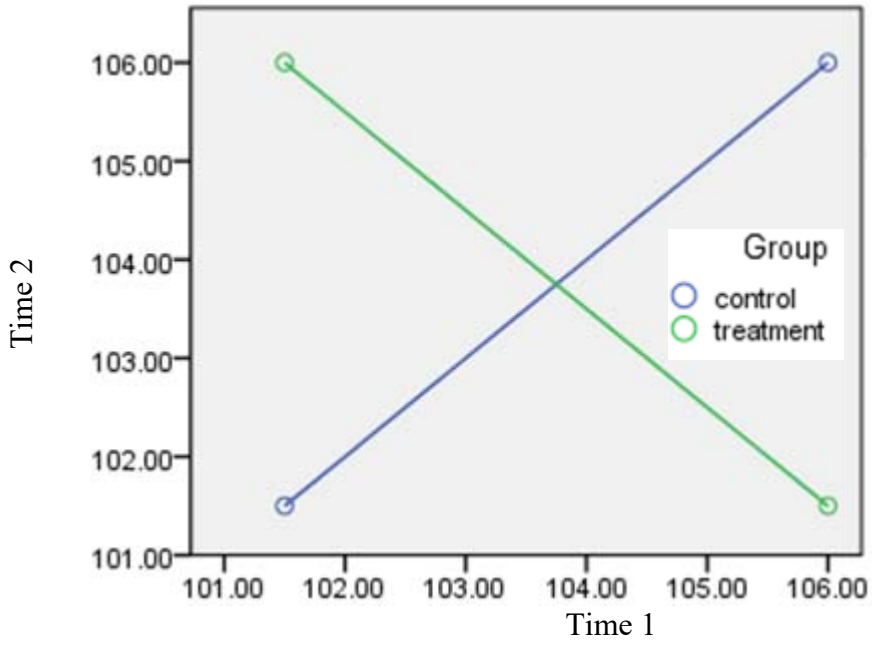
$$T_2 = B_0 + \boldsymbol{B_1} T_1 + B_2 \boldsymbol{G} + \varepsilon \quad (2)$$

The third experiment is conducted in two phases.

Experiment 3A: … with Symmetrical Slopes

The first phase of the third experiment assumes random participant allocation; $r_{(t1,t2)}$ is varied in a symmetrical linear fashion following a pattern of $-r$ / $+r$ for groups 1 and 2 respectively (Figure 1); assuming values of -.1 / .1, -.3 / ,3, -.5 / .5, -.7 / .7, and -.9 / .9. Sample size is fixed to 40 (20 per group); effect size is varied in increments of d =.1 by increasing the group 2 Time 2 mean by 1.5, from $d = 0$ to $d = 1.2$.

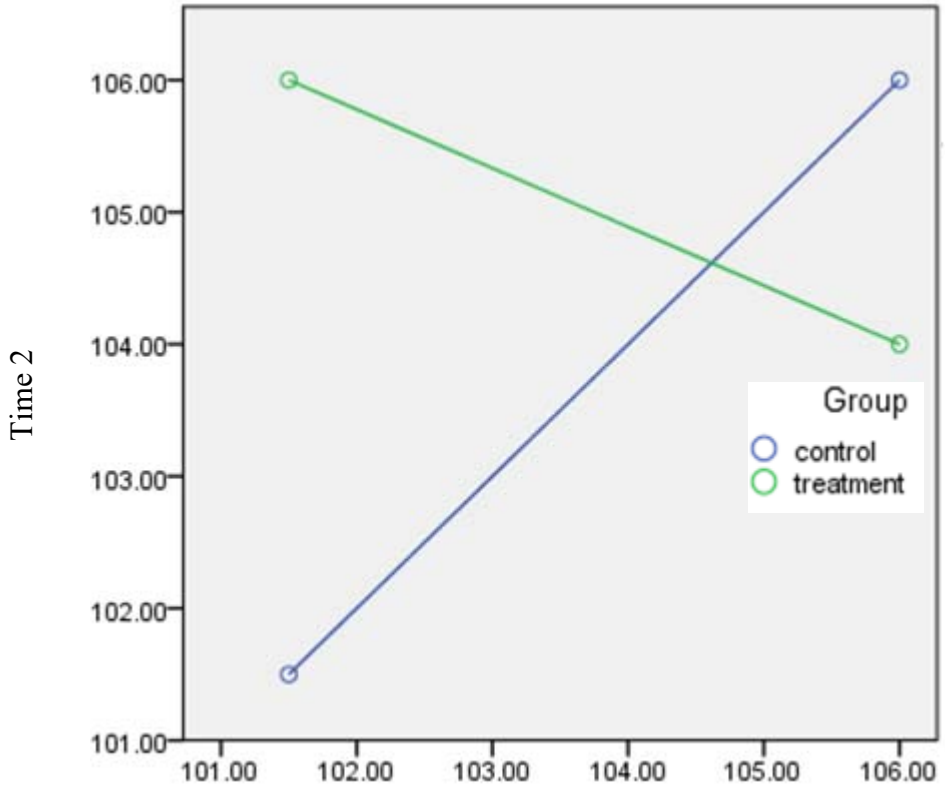
Figure 1: Symmetrical $r_{(t1,t2)}$ as a Function of Group



Experiment 3B: … with Asymmetrical Slopes

The second phase of the third experiment assumes random participant allocation; $r_{(t1,t2)}$ is varied in an asymmetrical fashion (Figure 2) following a pattern, assuming values of .3/.5, .1/.7, -.3/.9, -.9/.3, and -.6/.6 (a symmetrical value used for its absolute difference value for comparison). These values were chosen to compare the impact of a symmetrical or asymmetrical difference in values of $r_{(t1,t2)}$, when their absolute difference is the same (with absolute differences of .2, .6, and 1.2); these magnitude hold no significant meaning and were chosen by convenience to cover a wide range of possibilities. Sample size is fixed to 40 (20 per group); effect size is varied d =.1 by increasing the group 2 Time 2 mean by 1.5, from d = 0 to d = 1.2.

Figure 2: Asymmetrical $r_{(t1,t2)}$ as a Function of Group



Time 1

Experiment 4: Exploring the Effect of Large Systematic Baseline Differences

The goal of the fourth experiment is to test the effect of systematic group differences at Time 1. Sampling error is already implicitly accounted for by the nature of the Monte Carlo simulations with random data.

Since the goal of the ANCOVA and the difference score tests is to account for the relative starting point of a participant, this should translate to the group as well. I hypothesize that the ANCOVA will maintain its superiority in properly rejecting the null hypothesis.

In terms of equation 2, the large systematic baseline difference means that the origin (B_0) term would interact with the grouping variable (G). This is again not modeled in either the ANCOVA or RMANOVA equation, and so it is of interest to see where it has the largest impact.

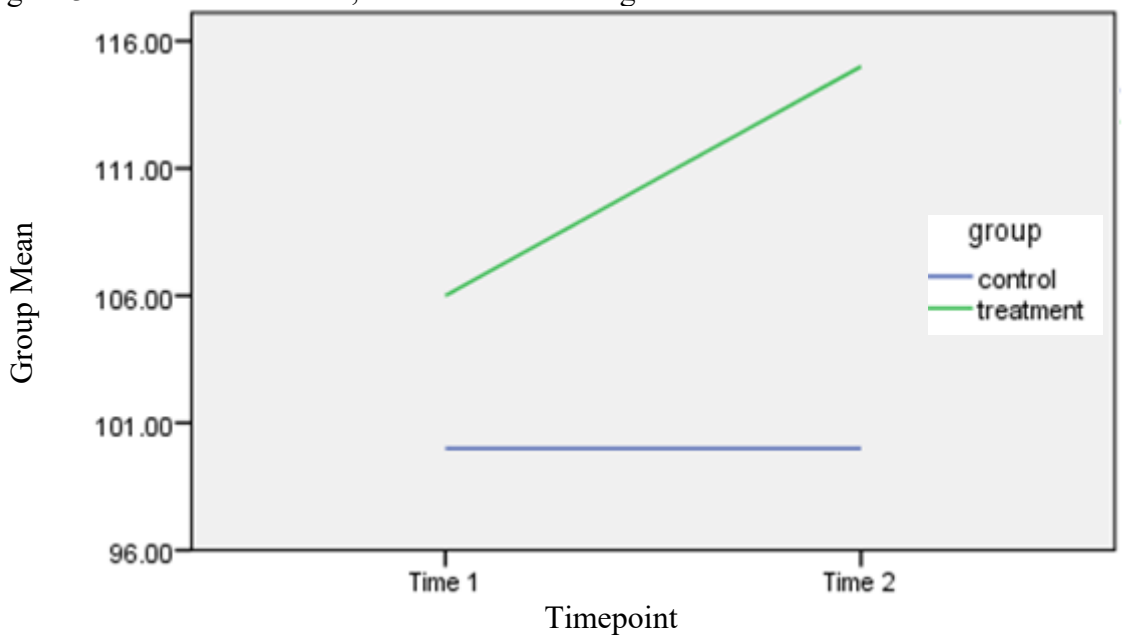
$$\mathrm{T}_2 = \boldsymbol{B_0} + B_1\mathrm{T}_1 + B_2\boldsymbol{G} + \varepsilon \qquad (2)$$

The fourth experiment is conducted in two phases. Both phases of the fourth experiment assume random participant allocation within two groups that are from different pre-existing groups. This is not a case of group assignment being done based on Time 1 scores.

Experiment 4A: … when the Treatment Group Starts Higher

$r_{(t1,t2)}$ is fixed to 0.5; n is fixed to 40 (20 per group); effect size is varied by increments of $d = .1$ by increasing the group 2 Time 2 mean by 1.5, from $d = 0$ to $d = 1.2$. This is repeated while assigning a higher Time 1 mean to group 2 (the treatment group) from $d = 0$ to $d = 0.9$ by increments of 0.1. This is illustrated in Figure 3.

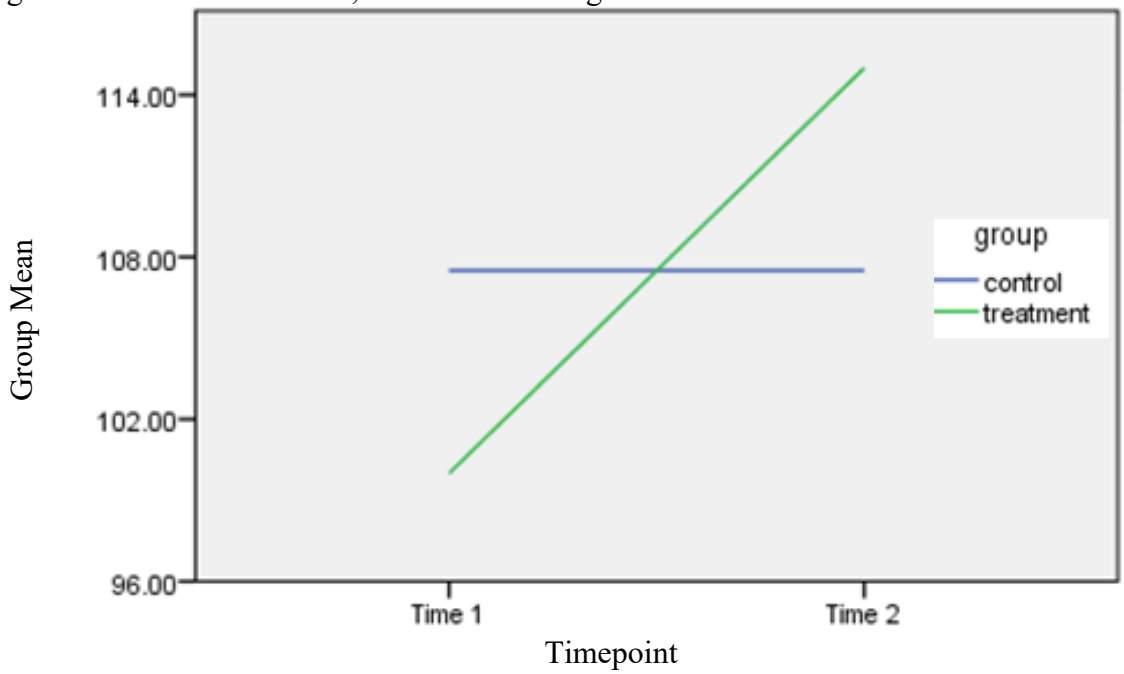
Figure 3 Baseline Difference; Treatment Starts Higher



Experiment 4B: … when the Control Group Starts Higher

Having a higher control group is of interest because this means that the two lines will cross as the treatment groups' effect grows. This makes it possible to observe the effect of a baseline imbalance scenario in which there is clearly a score growth, but a net nil difference between both groups in the end. This is illustrated by Figure 4.

Figure 4 Baseline Difference; Control Starts Higher



$r_{(t1,t2)}$ is fixed to 0.5; n is fixed to 40 (20 per group); effect size is varied by increments of $d = .1$ by increasing the group 2 Time 2 mean by 1.5, from $d = 0$ to $d = 1.2$. This is repeated while assigning a higher Time 1 mean to group 1 (this being the control group, this same higher mean is assigned to its Time 2 score as well) from $d = 0$ to $d = 0.9$ by increments of 0.1.

Experiment 5: The Impact of Missing Data

The goal of the fifth experiment is to observe the effect of randomly missing data on the respective power and Type I error of ANCOVA and difference score analyses within a very limited capacity.

The scenarios concerning missing data will be tested by using unbalanced groups. Given that missing data often follows patterns, especially when considering attrition

rates, modelling these patterns is well beyond the scope of this work. Rather, my goal is to simply assess the relative power of the respective methods. For these reasons, the missing data scenarios tested will be limited to randomly missing cases with listwise deletion. Given that randomness is explicitly specified, this is equivalent to simply having unbalanced group sizes. I will not be looking at other adjustment methods such as mean substitution, interpolation, or imputation. Using mean substitution would amount to testing the RMANOVA with data generated using that linear model, and using interpolation would amount to testing the ANCOVA with data generated using that linear model, *and would require a time series*. It seems to be of little interest to test such a tautology. Ultimately, testing missingness in such a restricted fashion will still provide a clear and definitive answer on whether missing data in and of itself is a greater issue for one method or another; whether the reasons for this missing information has a relational link with other aspects of the model is a different question.

I hypothesize that the ANCOVA will be less affected by this, given my prior hypothesis that its sample size requirements are smaller for a given power level, when compared to the RMANOVA. However, when it comes to the effect of having more participants missing in the control or treatment group, any further hypothesis would be speculative.

Experiment 5A: Participants Missing in the Control Group

The first part of the fifth experiment assumes random participant allocation; $r_{(t1,t2)}$ is varied from 0 to 1 in increments of .1; effect size is varied in increments of d =.1 by increasing the group 2 Time 2 mean by 1.5, from d = 0 to d = 1.2. Missing data in the

form of unbalanced groups is done by fixing the control group size to $n = 20$, while the test group size varies from $n = 21$ to $n = 30$ in increments of 1. This will cover attrition rates of up to 33%, which is considered normal in repeated studies with human volunteers.

Experiment 5B: Participants Missing in the Treatment Group

The second part of the fifth experiment assumes random participant allocation; $r_{(t1,t2)}$ is varied from 0 to 1 in increments of .1; effect size is varied in increments of $d = .1$ by increasing the group 2 Time 2 mean by 1.5, from $d = 0$ to $d = 1.2$. Missing data in the form of unbalanced groups is done by fixing the treatment group size to $n = 20$, while the control group size varies from $n = 21$ to $n = 30$ in increments of 1. This will cover attrition rates of up to 33%.

Chapter 4- Results

The results are presented in a way which allows for a discussion that covers the topic of the hypotheses laid out earlier.

The analysis of the results is done by collecting the observed power values for each test, and that result is converted into a dummy coded value. If the test was statistically significant at $\alpha = .05$, that is that $p < .05$, then the test gets a flag of 1. Otherwise, it is coded as 0. For the simple effects, there is the added caveat that if the interaction term was coded 0, the simple effect results are ignored and immediately coded as 0. This creates matrices of counts that allow for the comparison of total rejections of the null hypothesis, as well as an iteration by iteration comparison between the methods.

For the analysis of the RMANOVA results, a type III sum of square is used. The RMANOVA's Time 1 vs. Time 2 simple effect comparisons for each group uses the interaction term's error term as explained in Chapter 2. The ANCOVA was analyzed using the regression equation in equation 2.

$$\mathrm{T}_2 = B_0 + B_1\mathrm{T}_1 + B_2G + \varepsilon \qquad (2)$$

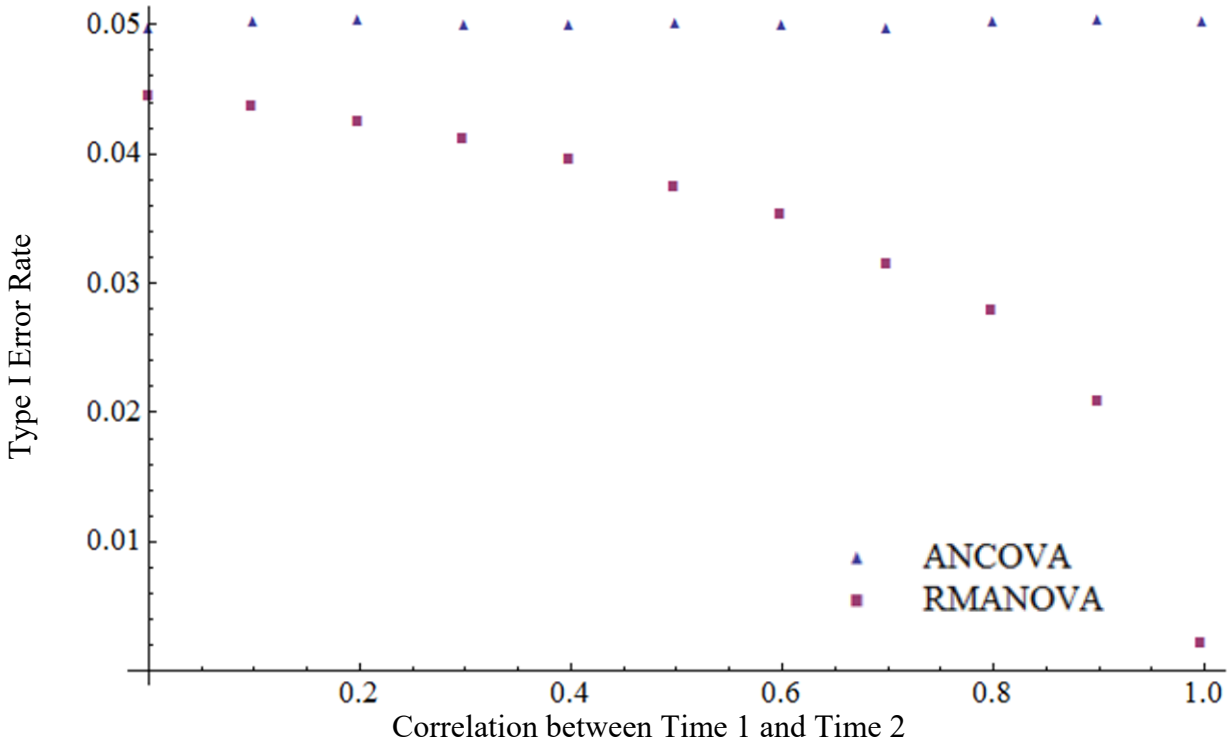
Experiment 1 (small sample): Varying the Effect Size Between Times; *n* = 20

Type I Error

The first element of interest is the rate of Type I errors. In Figure 5, we can see that the ANCOVA remains nearly perfectly in line with the projected Type I error rate. In fact, the largest deviation is when the correlation between Time 1 and Time 2 tends

towards unity, with a deviation of 0.000369 from 0.05; that is to say 369 supplementary Type I errors out of 1 000 000 attempts. On the other hand, the RMANOVA was overly conservative. The difference could be seen as marginal, especially when the correlation is small, but the disparity grows to the point of nearly eliminating Type I errors. When the correlation is nearly perfect, the Type I error rate is of 0.002186, deviating by a substantial -0.047814 from 0.05.

Figure 5 Type I Error Rates as a Function of $r_{(t1,t2)}$ when $n = 20$

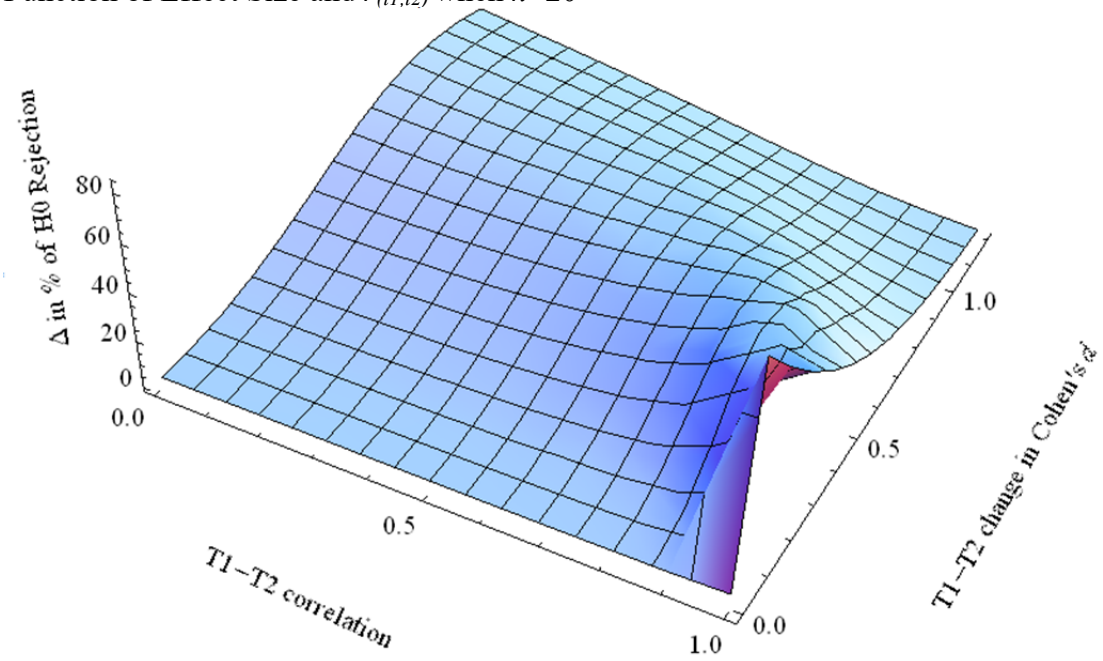


Net H_0 Rejection Rate Difference

The second element of interest is the power achieved by either test. Figure 6 is a plane representing the difference of successful rejection of the null hypothesis between the ANCOVA and the RMANOVA's contingent simple effect. The RMANOVA does

not, at any point, outperform the ANCOVA. This can be seen by the fact that the plane is always above zero. There is an anomalous spike when the correlation between Time 1 and Time 2 is nearing unity and the effect size is negligible ($d = 0.1$), where the ANCOVA reported a statistically significant difference on all iterations, except for two iterations out of 12 million. When a perfect value of $r_{(t1,t2)} = 1$ was used (rather than 0.999 as is the case here), the ANCOVA failed to compute as the variance became null, and it is likely safe to assume that the sensible values to observe are the ones at $r_{(t1,t2)} = 0.9$ and below. The value was still included given the importance of the equivalences of the models when $r = 1$, and the topic is further explored in the discussion. The peak difference between the two tests is a gap of 44% supplementary correct rejections of the null hypothesis by the ANCOVA at $r_{(t1,t2)} = 0.9$ and an effect size of $d = 0.4$. As the correlation between the two time points is reduced, the peak difference transitions steadily to be at a large effect size of $d = 1$ when $r_{(t1,t2)} = 0.1$, and remains quite large, with the total success rate of the ANCOVA being 40 percentage points higher. This difference slowly shrinks to 36 percentage points when $r_{(t1,t2)} = 0.8$ before rising again when $r_{(t1,t2)} = 0.9$.

Figure 6 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and $r_{(t1,t2)}$ when n=20



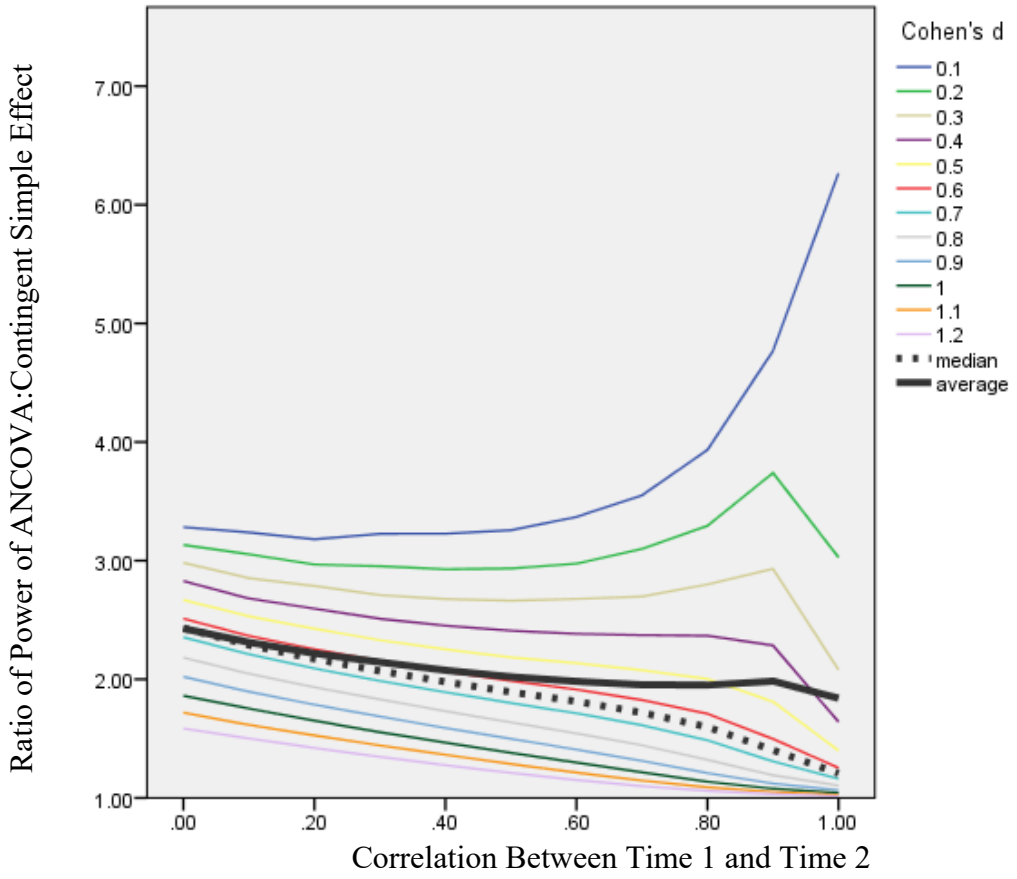
There is a clear trend in the data whereby the RMANOVA's power is generally approximately equivalent to that of the ANCOVA with a lag of approximately 3 steps in effect size (d = 0.3).

H_0 Rejection Rate Ratio

To address the question of whether the relative advantage of the ANCOVA diminishes as the correlation between Time 1 and Time 2 increase, the ratio of the ANCOVA's power compared to the RMANOVA's contingent simple effect power was calculated for each of the 110 combinations tested (this excludes the d = 0 scenario of Type I error), spanning 110 million iterations. As we can see in Figure 7, the ANCOVA's relative advantage does generally decrease as $r_{(t1,t2)}$ increases. The clear exception to this is when there is a combination of a very high correlation and a very low effect size. The increase in the ANCOVA's relative advantage affects a wider range of $r_{(t1,t2)}$ values as

Cohen's d decreases; but the majority of this is localized within values of *d* smaller than 0.3, and of $r_{(t1,t2)}$ greater or equal to 0.7. When we look at the mean and the median, the pattern becomes much more consistent, but even then the average sees a plateau at $r_{(t1,t2)}$ = 0.8, with a slight spike at $r_{(t1,t2)}$ = 0.9. The grand average for a medium correlation of $r_{(t1,t2)}$ = 0.5 is 1.98, which indicates that the ANCOVA has about twice the power of a RMANOVA.

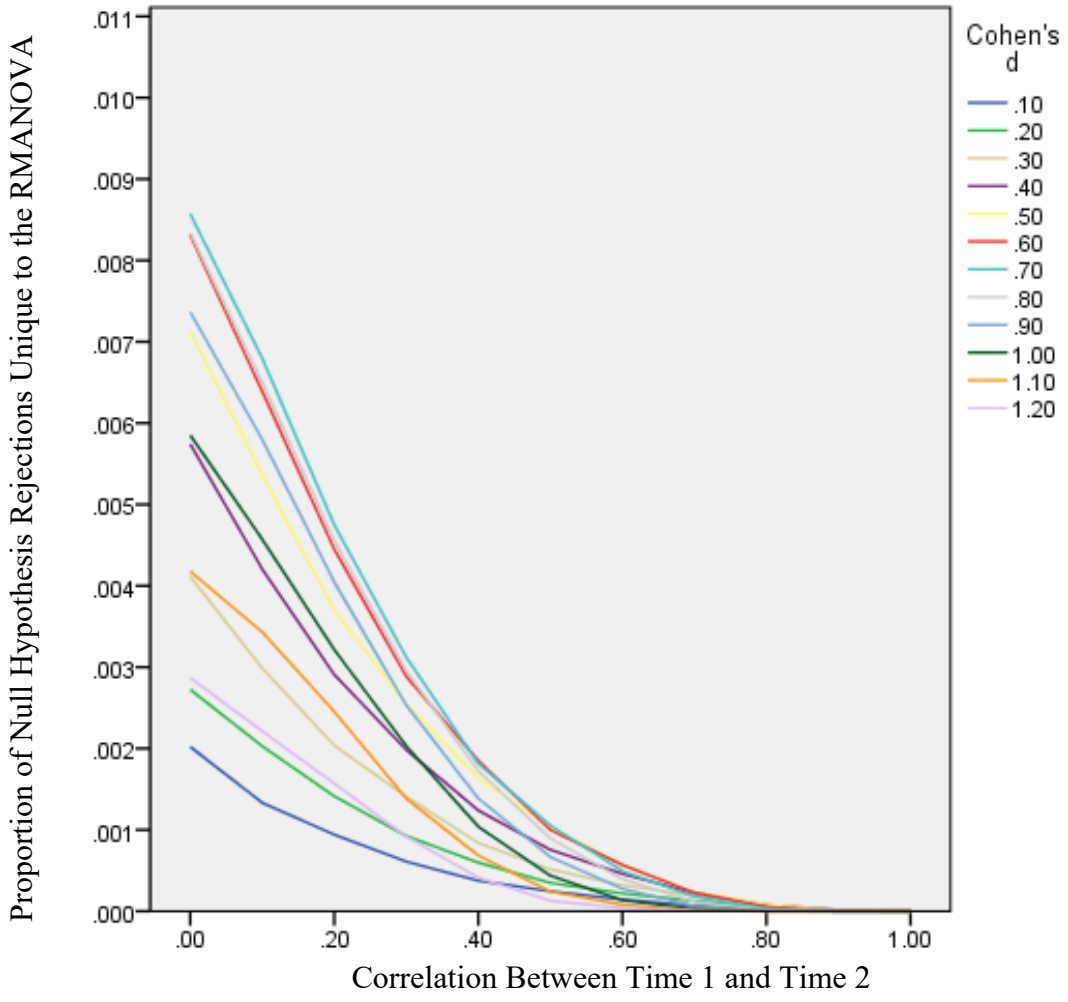
Figure 7 Ratio of Power Between ANCOVA and RMANOVA as a Function of $r_{(t1,t2)}$



Unique H_0 Rejections

The last element of interest from this experiment is the number of unique null hypothesis rejections. Given that the RMANOVA has far less success at rejecting the null hypothesis than the ANCOVA, it is the RMANOVAs' unique rejections that are of interest to evaluate the level of overlap between the two methods. This information is illustrated in Figure 8. The peak level of unique rejections occurs with small correlations between Time 1 and Time 2 at medium to large effect sizes of $d = 0.4$ to $d = 1$, approaching a rate of 1% and never surpasses 0.1% when that value reaches $r_{(t1,t2)} = 0.5$.

Figure 8 Proportion of Unique RMANOVA Rejections of H_0 as a Function of $r_{(t1,t2)}$



The ANCOVA far more frequently rejects the null hypothesis where the RMANOVA does not. Figure 9 illustrates those unique rejections of the null hypothesis by the ANCOVA. Figure 10 displays the unique rejections of the null hypothesis by the RMANOVA, but on the same scale as the ANCOVA's unique rejections.

Figure 9 Proportion of Unique ANCOVA Rejections of H_0 as a Function of $r_{(t1,t2)}$

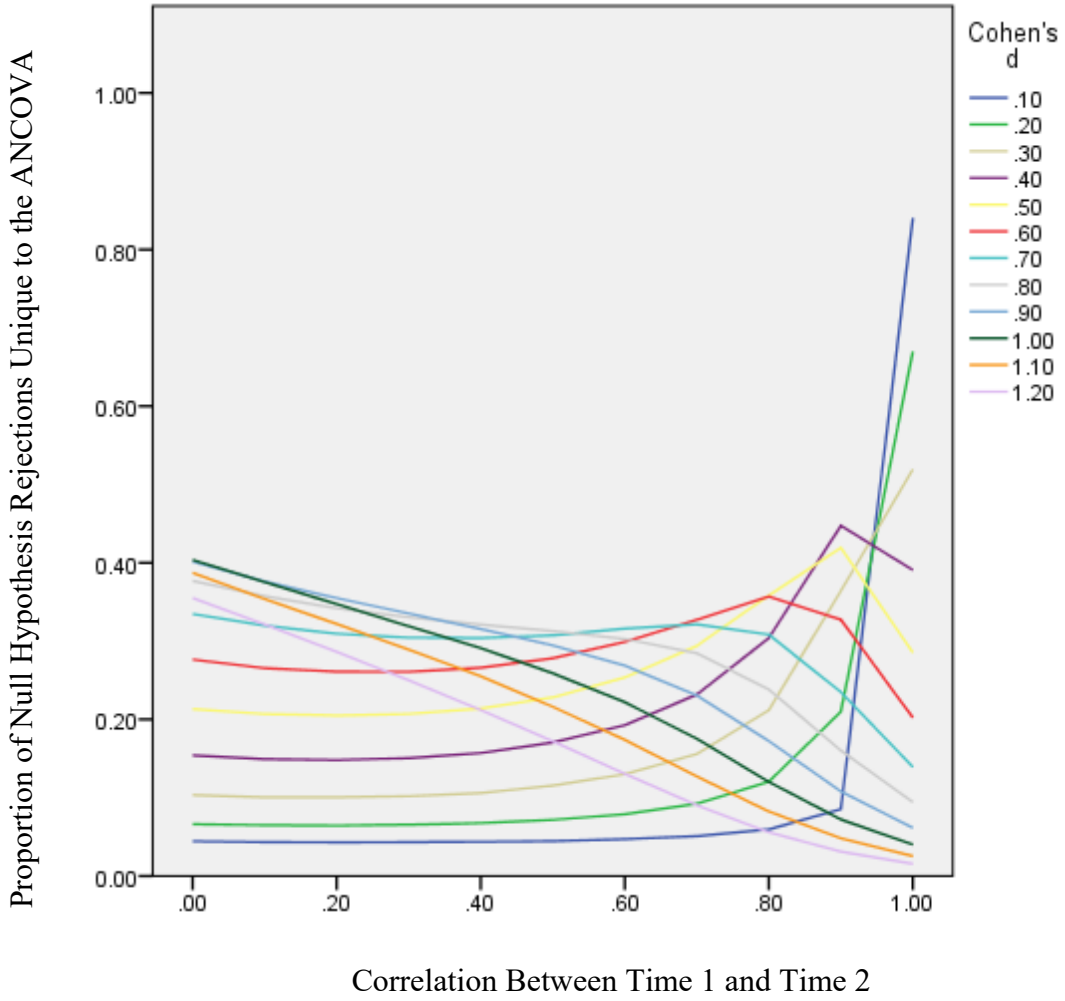
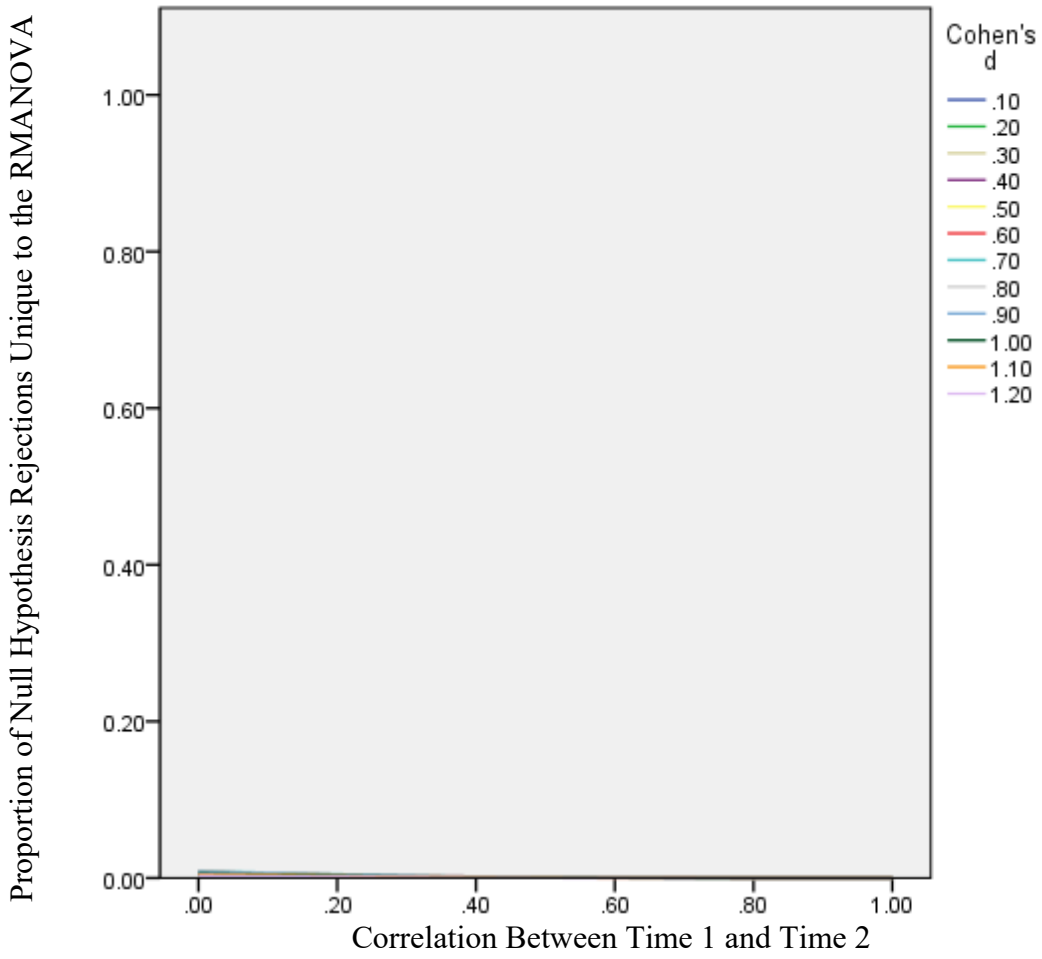


Figure 10 Proportion of Unique RMANOVA Rejections of H_0 (Figure 8) as a Function of $r_{(t1,t2)}$ on the scale of Figure 9

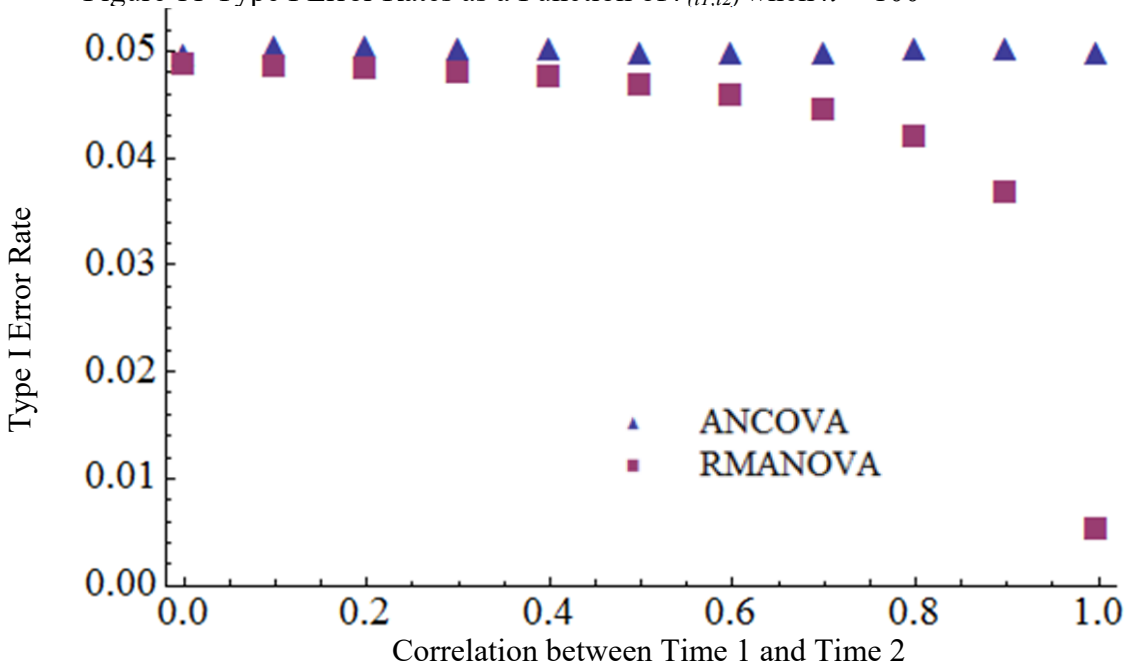


Experiment 1 (large sample): Varying the Effect Size Between Times; *n* = 100

Type I Error

We can see in Figure 11 that as with small samples, the ANCOVA remains nearly perfectly in line with the projected Type I error rate. The RMANOVA's Type I error rate goes down as the correlation goes up, down to 0.53% when the $r_{(t1,t2)} = 0.999$. However, at the $r_{(t1,t2)} = 0.9$, the RMANOVA's Type I error rate is still of 3.7%.

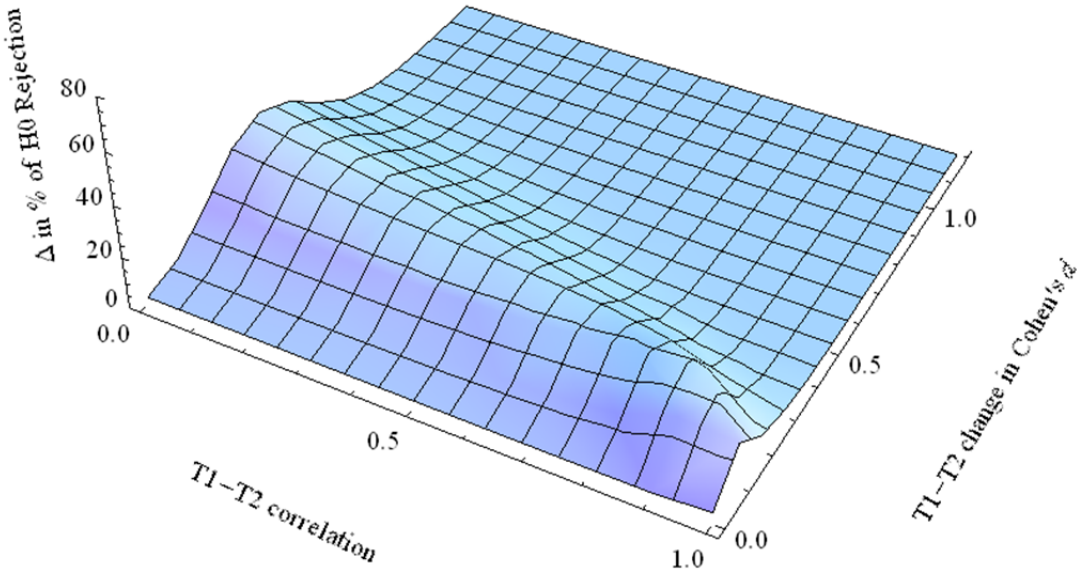
Figure 11 Type I Error Rates as a Function of $r_{(t1,t2)}$ when n = 100



Net H_0 Rejection Rate Difference

In Figure 12, the peak of the wave-like form has shifted down to below d = 0.5 compared to the n = 100 scenario. The peak difference is around 33%, and declines as the conditions become less difficult. The difference between the two tests becomes nil with high effect sizes and medium correlations between the time points. However, at low effect sizes, the difference remains as high as 19.4% when $r_{(t1,t2)}$ = 0.7, before bouncing back up to 24.1% as $r_{(t1,t2)}$ increases to 0.9.

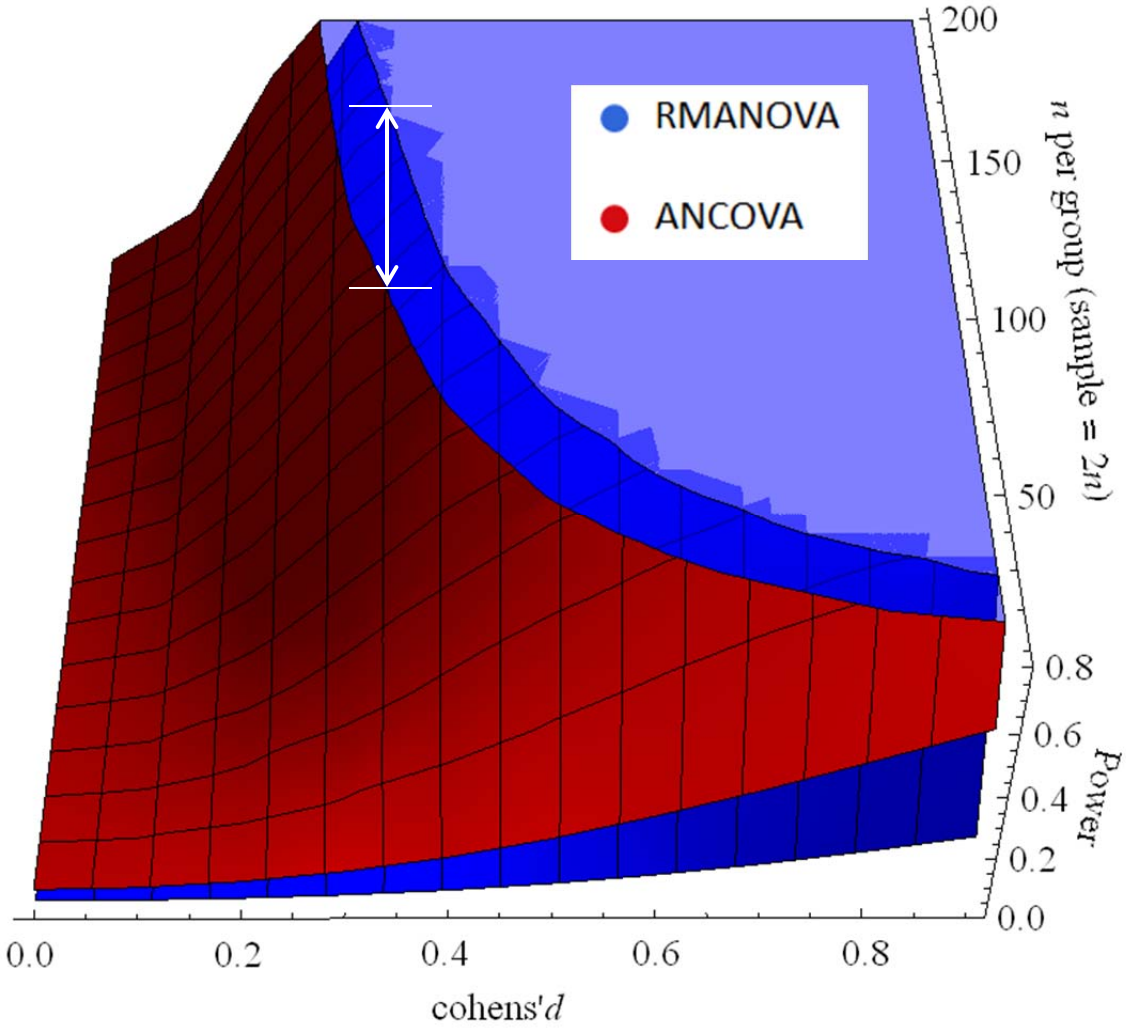
Figure 12 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and $r_{(t1,t2)}$ when n=100



Experiment 2: Determining the Sample Size Required to Obtain .80 Power

Figure 13 illustrates the power curve as group size (n) and effect size (d) increase. The figure has a threshold at a power of 0.80 so that the 80% power delineation can clearly be seen. The ANCOVA always achieves a power of 0.80 before the RMANOVA, regardless of the effect size or the sample size. With a large sample size and smaller effect size, or large effect size and smaller sample size, both tests rapidly reach the desired power, but the ANCOVA is unequivocally more efficient.

Figure 13 Reaching a Power of 0.80 as a Function of *n* and *d*. The arrow corresponds to the sample size requirement to reach a power of 0.80 between the ANCOVA and RMANOVA when *d* = 0.37.



When both tests reach the requisite threshold of 0.8 for a given effect size (*d*), the sample size requirement for the RMANOVA is slightly more than 1.5 times larger than that of the ANCOVA (mean ratio difference of 1.56, median of 1.53). This was calculated by interpolating the sample size required to reach an actual power of 0.8 from the nearest mean value produced by the simulations.

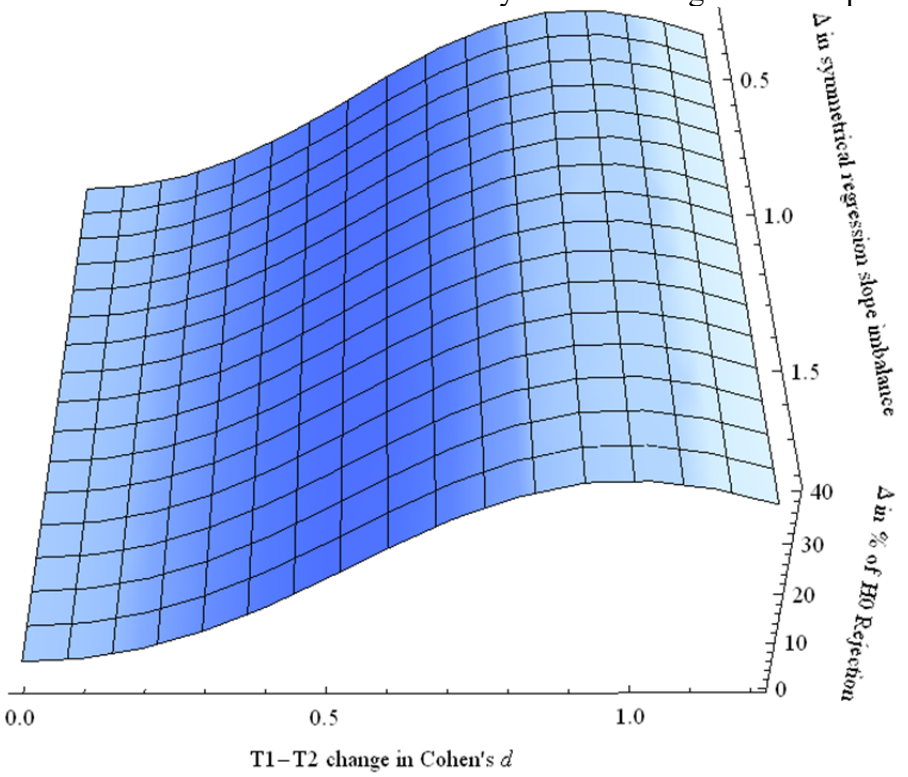
Experiment 3: Violating the Homogeneity of the Regression Slopes Assumption

As the correlation between Time 1 and Time 2 become different for the two groups, it can be said that there is an interaction between the covariate and the group, and therefore the homogeneity of the regression slopes assumption has not been respected.

Experiment 3A… with Symmetrical Slopes

When I speak of a symmetrical slope imbalance, I am referring to values that are equal in an absolute form (-1/+1, -3/+3, etc.). As can clearly be seen in Figure 14, changes in the magnitude of a symmetrical difference (i.e. -1/+1 vs -9/+9) have absolutely no impact on the differential power profile, and as such, no impact on the power profile of either method.

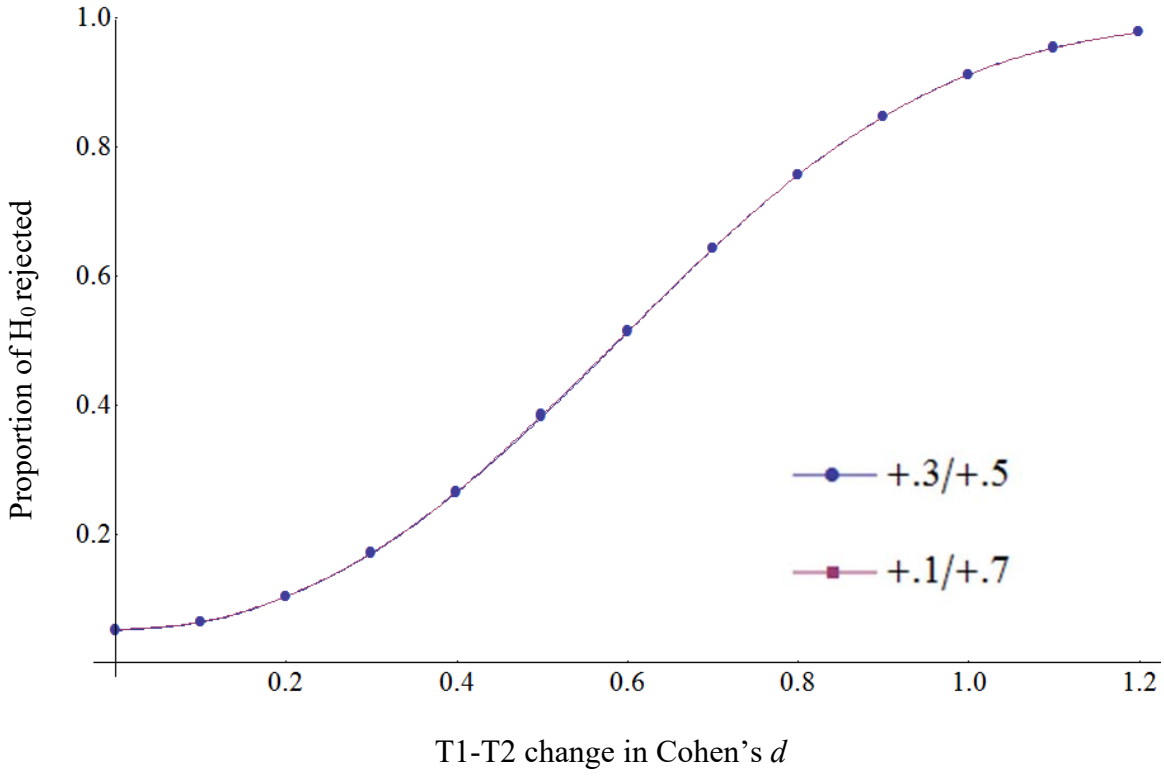
Figure 14 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and Level of Symmetrical Regression Slopes Imbalance



Experiment 3B … with Asymmetrical Slopes

Exploring combinations of asymmetrical slopes sheds some light on why the symmetrical slopes give the same results. By looking at the impact of two pairs of asymmetrical heterogeneous slopes, +.3/+.5 and +.1/+.7, we can replicate this absence of a difference as shown in Figure 15. These values were chosen to illustrate a case where the difference between the two values are different (0.5 – 0.3 = 0.2 and 0.7 – 0.1 = 0.6), but the means are identical (in both cases, the average is of 0.4).

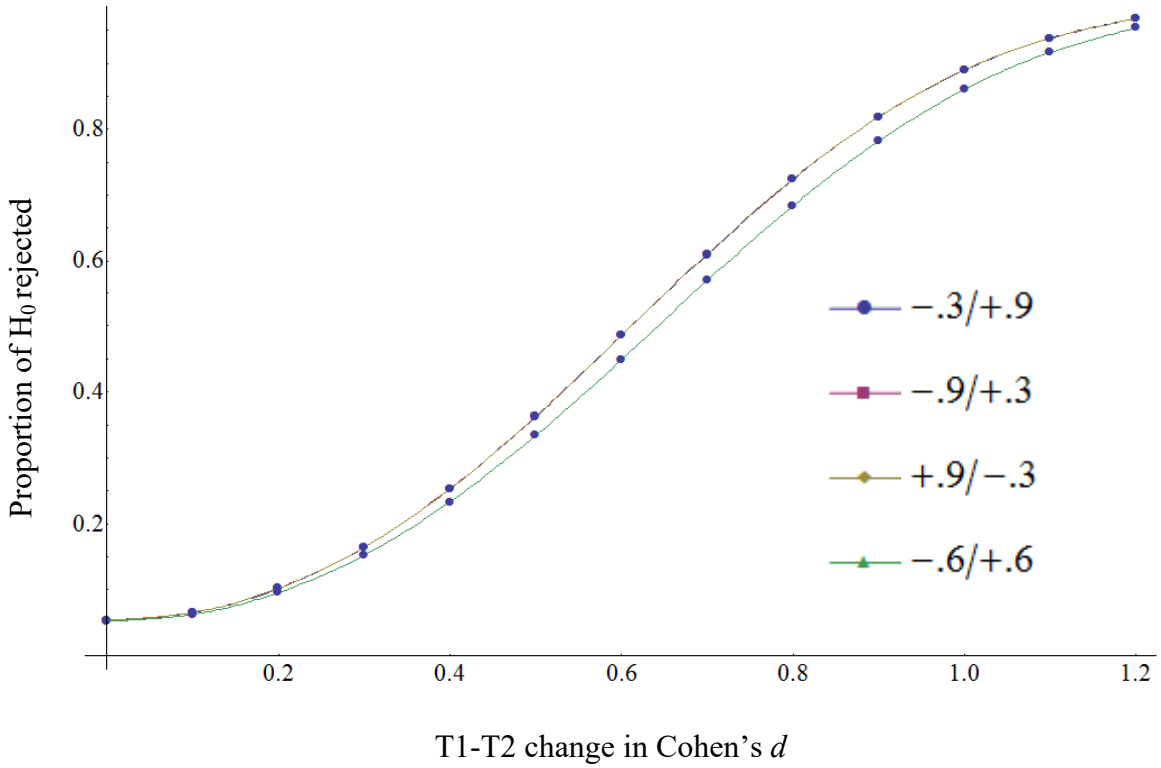
Figure 15 Rejection Rate of H0 as a Function of Treatment Effect Size for Cases where Average of Slopes is Equal



In Figure 16, there is a comparison of three variants of asymmetrical slope pairs with the same mean, as well as another pair which has a different mean than the first

three, but has the same differential between the two, to rule out the possibility that difference between the slopes plays a role in this pattern. The curves diverge as expected when the means are not the same.

Figure 16 Rejection Rate of H0 as a Function of Treatment Effect Size for Cases where Difference of Slopes is Equal



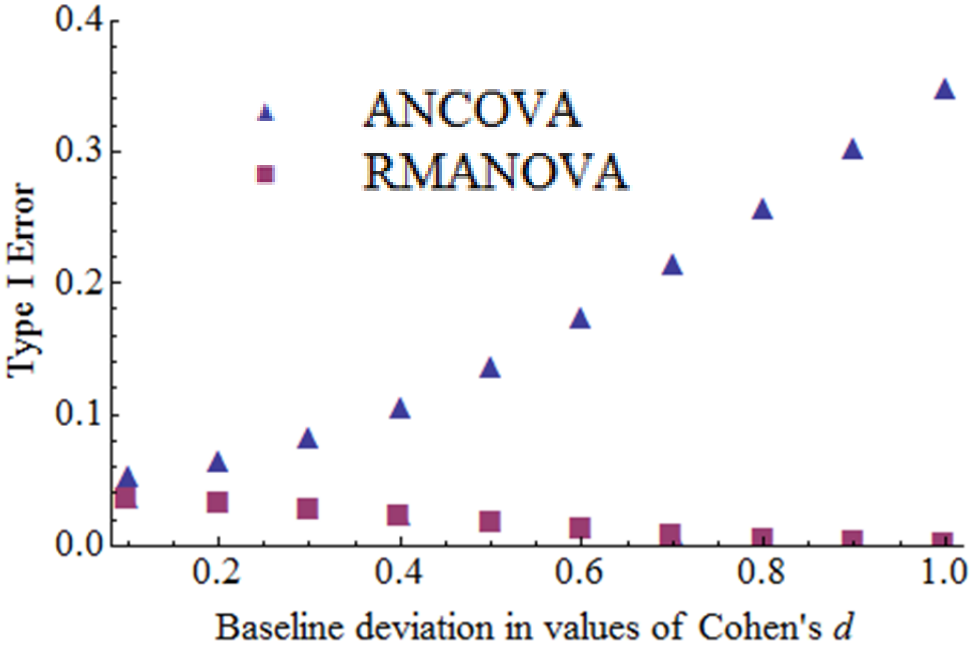
Experiment 4: Exploring the Effect of Large Systematic Baseline Differences

The impact of large systematic baseline differences was explored both in cases where the treatment group had systematically higher Time 1 scores, and where the control group had systematically higher Time 1 scores. In both cases, I looked at the impact on the power differential profile as well as Type I error. In this case, the RMANOVA's interaction term was used, as changing the baseline values introduces errors in the simple effects, as a difference is in fact present when it is analyzed in isolation from the other factors being controlled (that is to say the Time 1 data in this case).

Experiment 4A: … when the Treatment Group Starts Higher

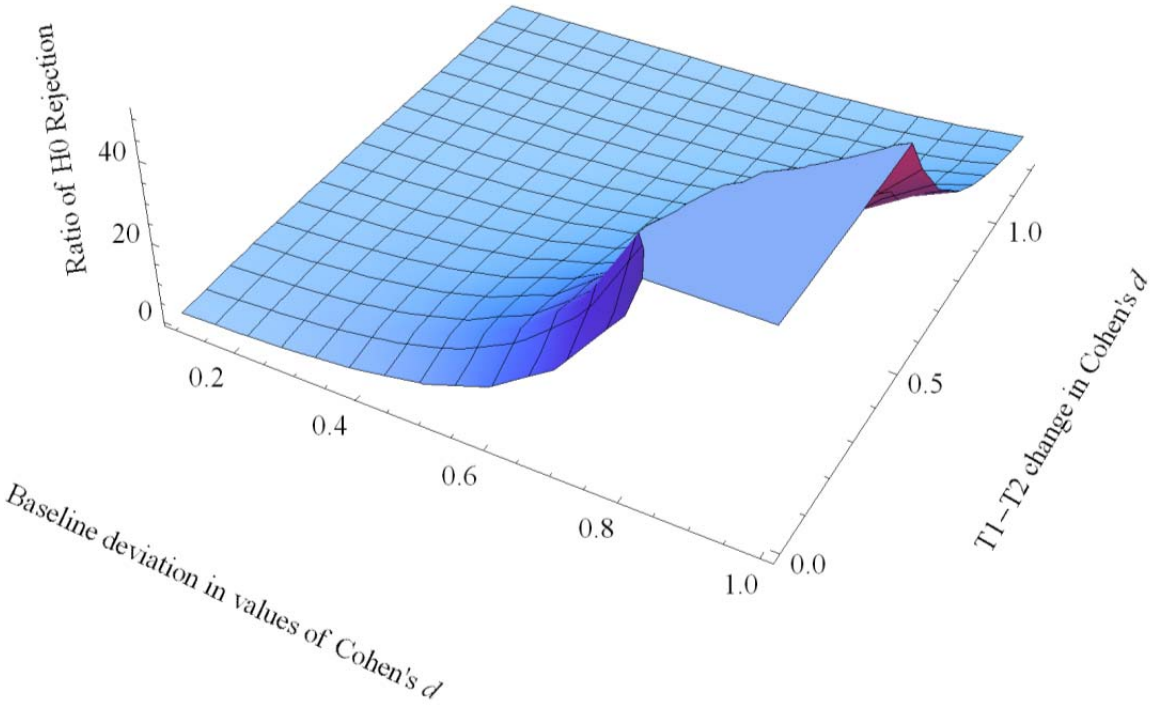
Figure 17 illustrates the change in baseline differences where the treatment group has a higher base mean than the control group. The ANCOVA's Type I error rate becomes significantly inflated, while the RMANOVA's drops. This inflation of the Type I error renders the analyses of further results irrelevant, as there would be no way to trust the outcome of a significance test when the Type I error is anywhere from 10% to 35% for medium and large effect sizes.

Figure 17 Type I Error Rates as a Function of $r_{(t1.t2)}$ when Treatment Group Starts Higher



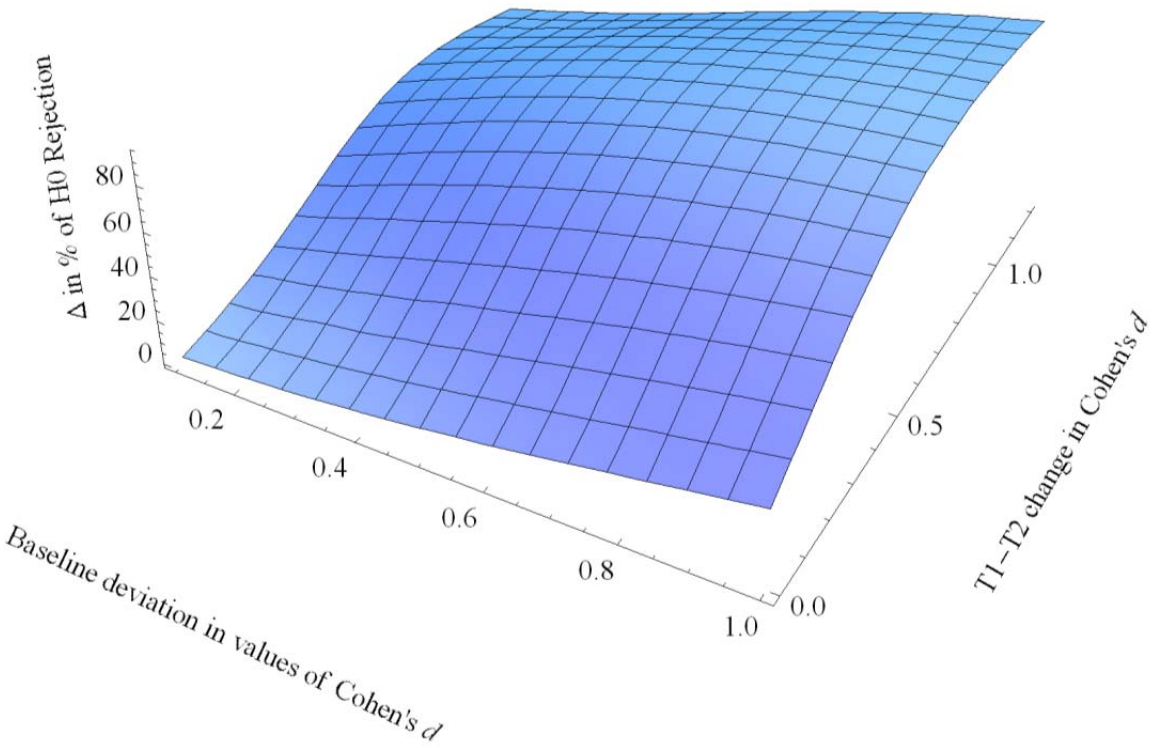
The ANCOVA unsurprisingly rejects far more tests given it's severely inflated Type I error rate. The ANCOVA rejects over 100 times the number of tests than the RMANOVA in low effect size and high baseline deviation situations (Figure 18).

Figure 18 Ratio of H0 Rejections as a Funciton of Effect Size and Level of Baseline Imbalance with Treatment Group Higher



As effect size and baseline difference are varied, we can see that the highest differences in overal number of tests rejected are not in the low effect size and high baseline deviation scenarios, but rather in the high effect size situations (Figure 19).

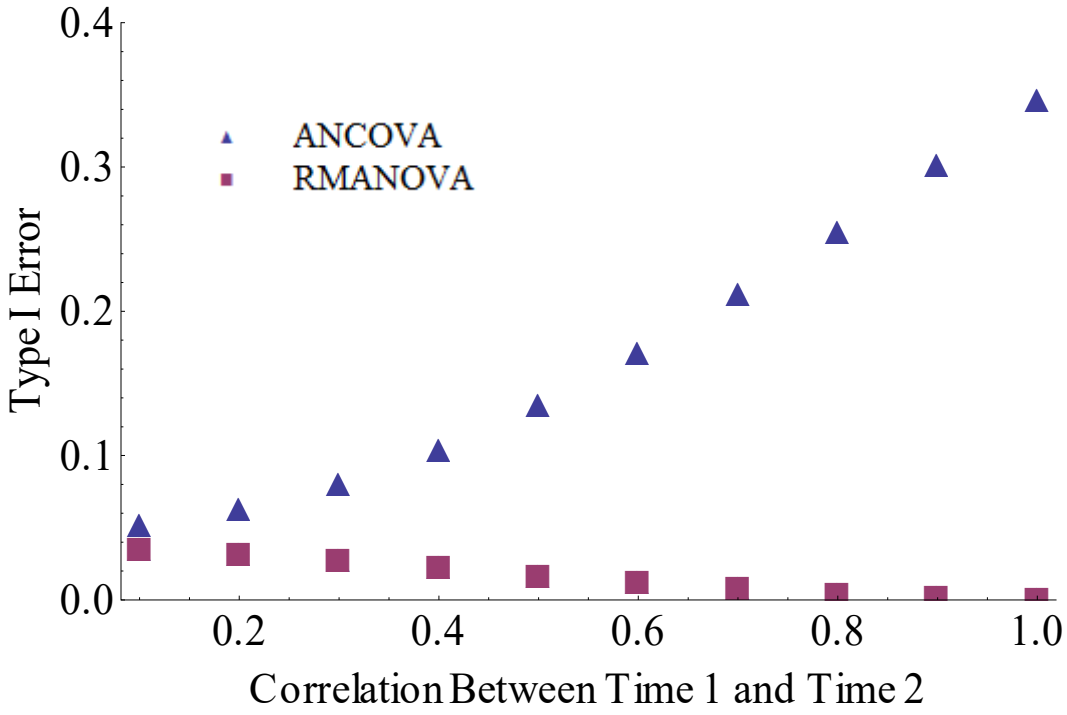
Figure 19 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and Level of Baseline Imbalance with Treatment Group Higher



Experiment 4B: … when the Control Group Starts Higher

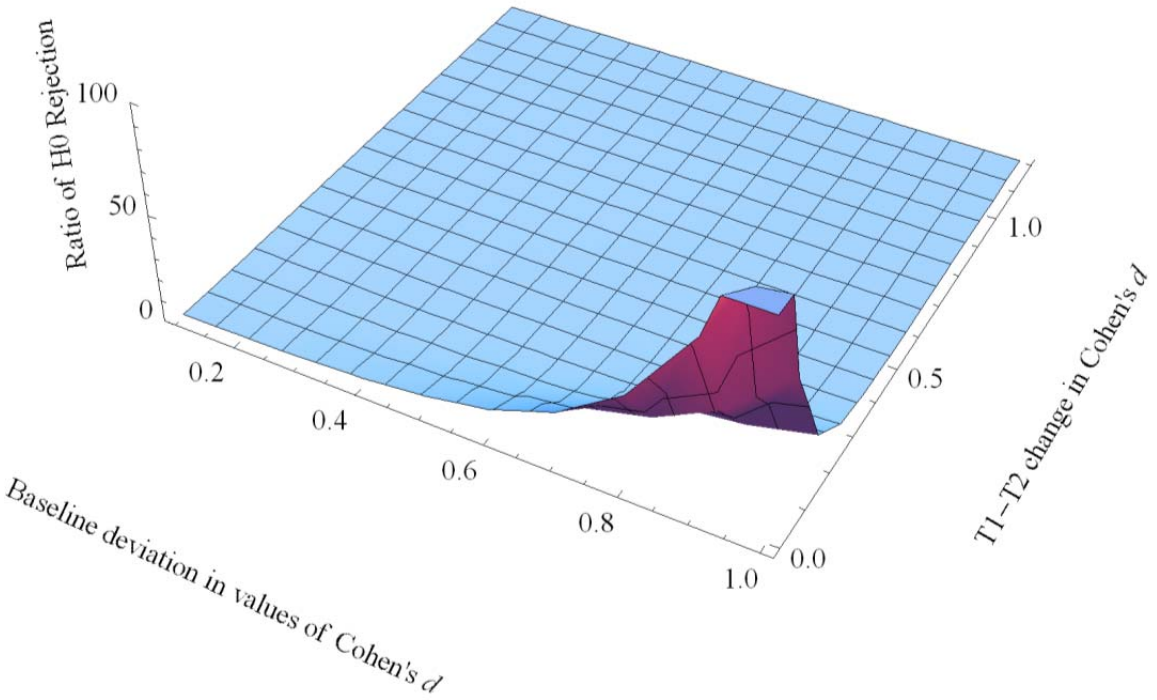
The same patterns can be seen in the case where the control group has a higher mean than the treatment group at the outset. Figure 20 illustrates that the change in baseline differences where the control group has a higher base mean than the treatment group has the same distorting effect on the ANCOVA's Type I error rate. Again, this renders the point of sucessful rejections of the null hypothesis moot.

Figure 20 Type I Error Rates as a Function of $r_{(t1,t2)}$ when Treatment Control Starts Higher



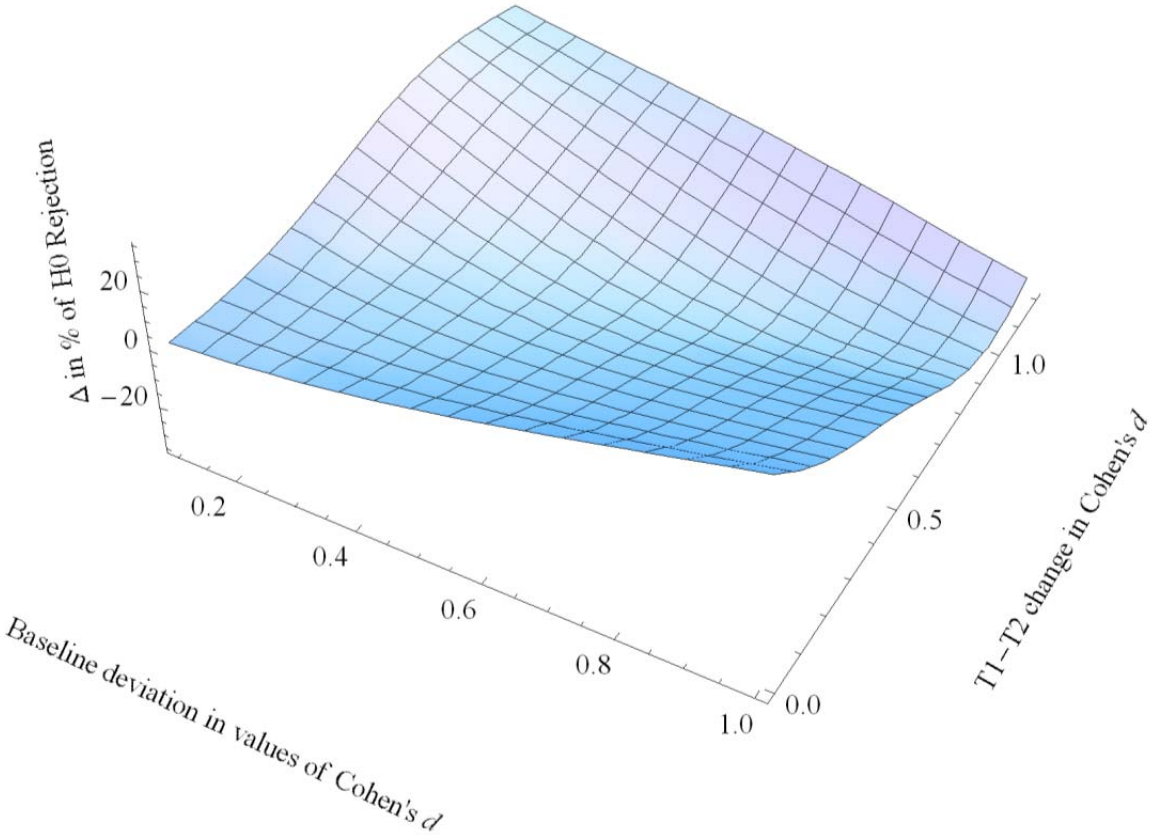
When the treatment group is relatively smaller to the control group, the ratio between the ANCOVA and the RMANOVA's rejection rates much more subdued, but still spikes in the hundreds when the baseline deviation is large and the effect size is small (Figure 21).

Figure 21 Ratio of H_0 Rejections as a Funciton of Effect Size and Level of Baseline Imbalance with Control Group Higher



The raw difference in the number of null hypoteses rejected peaks with low baseline deviation combined with high effect size, as well as when there is a high baseline deviation and a low effect size. However, the RMANOVA rejects more null hypotheses than the ANCOVA when the baseline deviation and the effect size are both high with a difference of -20% (Figure 22).

Figure 22 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and Level of Baseline Imbalance with Control Group Higher



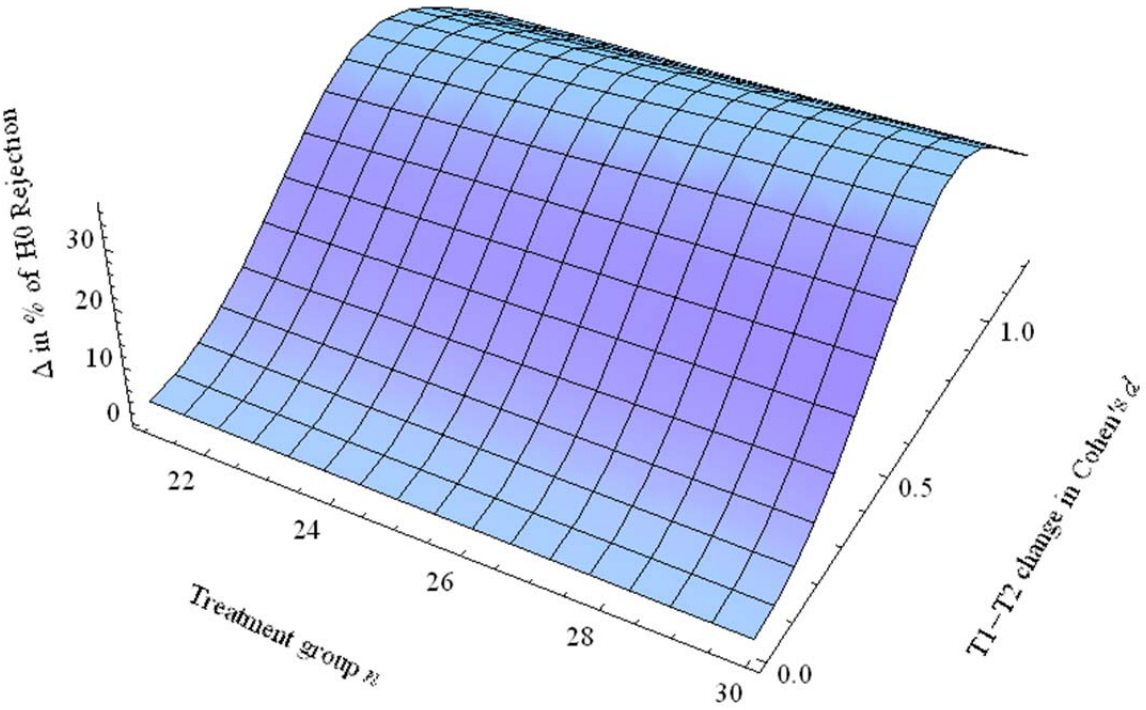
Experiment 5: The Impact of Missing Data

Experiment 5A: Participants Missing in the Control Group

When the group imbalance is to the disadvantage of the control group (the treatment group is larger), the graph looks fairly uniform (Figure 23); but there is a slow shift in relative advantage of the ANCOVA over the RMANOVA. In absolute values, the success difference keeps the same quadratic shape as with the other experiments. The peak difference is found at $d = 0.8$ and ranges from a difference of 31% fewer correct

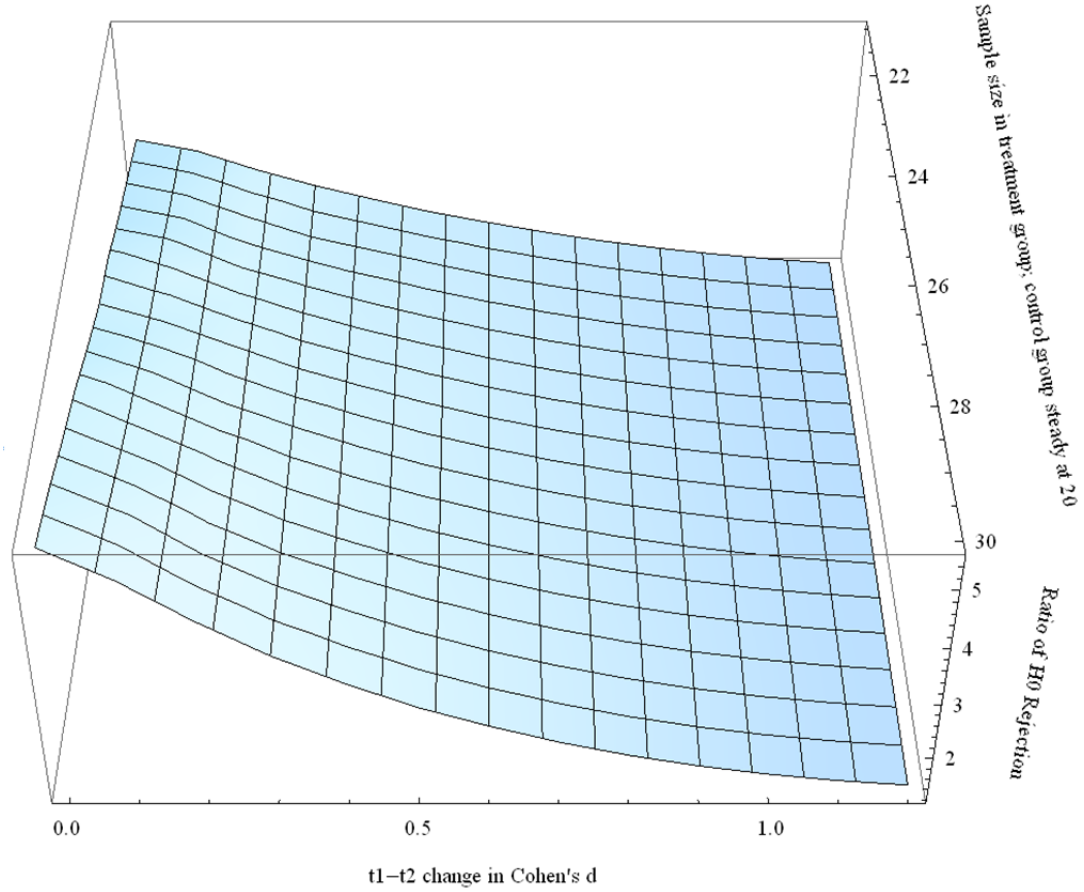
rejections of the null hypothesis for the RMANOA than for the ANCOVA when there is little group size difference, to a difference of 36% when the gap is largest.

Figure 23 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and Level of Listwise Deletion in Treatment Group



The success ratio displayed in Figure 24 clearly shows that the ratio is dependent on the group imbalance, as well as the effect size. That the effect size has an impact is of no surprise, as it was demonstrated earlier. The imbalance peaks when the treatment group is largest in relation to the control group. Upon closer inspection of the data, we can see that the ratio of ANCOVA to RMANOVA successes increases most when the group size disparity increases for small effect size samples. This amounts to an average shift of 0.136 in the disparity ratio per increase of the group size by 1 for effect sizes of d=0.3 or less. This decrease in the success ratio plateaus at $d = 1$.

Figure 24 Ratio of H_0 Rejections as a Funciton of Effect Size and Level of Listwise Deletion in Treatment Group



Experiment 5B: Participants Missing in the Treatment Group

When the missing data is found in the treatment group (the control group is larger), the peak difference is once again found at $d = 0.8$. However, the disparity is much smaller, peaking at 30% when the groups are nearly equal in size, and going down to 24% when the group size difference is greatest (Figure 25).

Figure 25 Difference of % of ANCOVA H_0 Rejection to RMANOVA H_0 Rejections as a Function of Effect Size and Level of Listwise Deletion in Control Group

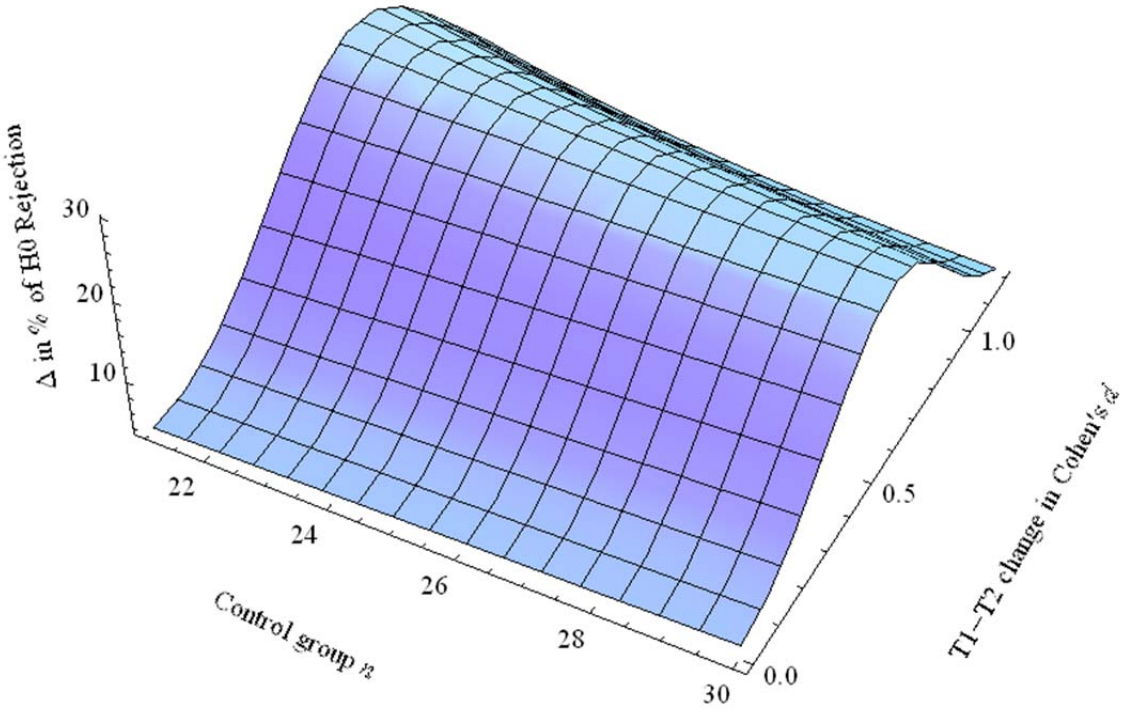
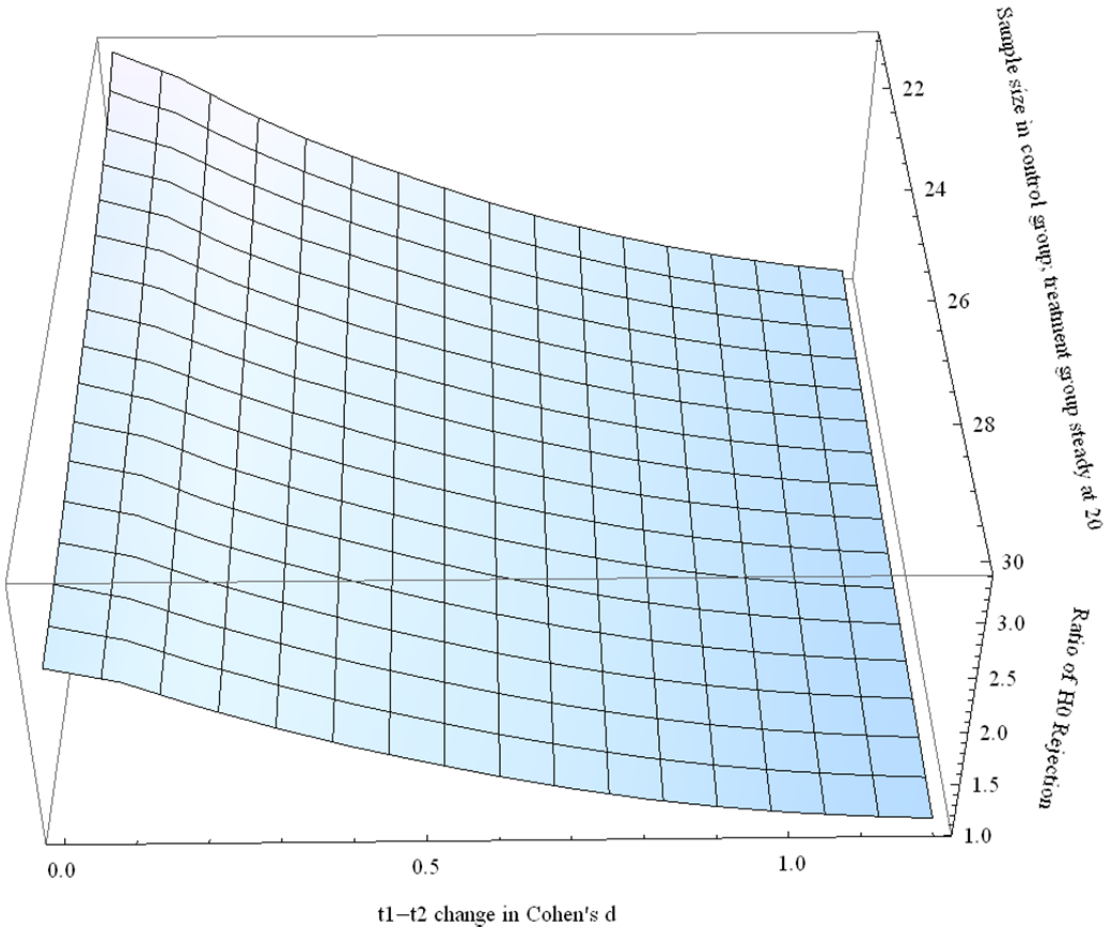


Figure 26 illustrates the shift in the success rate ratio. The ANCOVA remains in a favourable position, but loses part of its relative advantage as the disparity in group sizes increases. This is the mirror image of what was found in the prior experiment. This transition is much slower, with the peak shift being found at values of d = 0.3 and lower once again, but this time representing an average decrease of 0.07 in the multiplication factor for each increase of the group size disparity by 1.

Figure 26 Ratio of H_0 Rejections as a Funciton of Effect Size and Level of Listwise Deletion in Control Group



Difference between Interaction results and those of the Contingent Simple Effect

At all times the data from the interaction terms was collected. The difference in power between the RMANOVA's interaction term and the contingent simple effects was greatest when the test was in a condition of low power (small effect size, small relationship between Time 1 and Time 2 data). The differences are not negligible, but do follow a consistent pattern. More importantly, they never change the direction of the results. The best way in which I can summarize these results is in talking about the ratio

between the ratios of the ANCOVA : interaction null hypothesis rejections and the ANCOVA : contingent simple effect null hypothesis rejections.

When looking at the overall data a smaller sample size of 40, when the effect size was small ($d = 0.1$ to $d = 0.3$), the difference between the ANCOVA and the RMANOVA would have been half of what it was with the contingent simple effect, had the interaction term been used directly. That is to say, if the ANCOVA correctly rejected the null hypothesis three times more frequently than the RMANOVA's contingent simple effect, then the ANCOVA correctly rejected the null hypothesis 1.5 time more frequently than the RMANOVA's interaction. For medium effect sizes, this ratio changed to approximately 1.35. Using the average of the ANCOVA correctly rejecting the null hypothesis approximately 2 times more frequently than the contingent simple effect, then that value would change to approximately 1.48 time for the interaction term. Finally, for large effect sizes ($d = 0.7$ and up), this ratio dropped to around 1.055. This means that the average ratio of ANCOVA correctly rejecting the null hypothesis over the contingent simple effect would go down from its average of roughly 1.4 to approximately 1.33.

The ratio of these ratios remains fairly stable across all levels of the correlation between Time 1 and Time 2 for each value of d as long as sample size remains stable. All of these ratios are also strongly affect by other elements that add to power. For example, when the sample size is greater, this disparity shrinks much more rapidly. With a sample size of 200 instead of 40, that ratio of ratios is approximately 1.4 for small effect sizes, drops to under 1.02 for medium effect sizes, and under 1.00003 for larger effect sizes.

Chapter 5- Discussion

Experiment 1: Varying the Effect Size Between Times

There is a difference in the Type I error rate between the two tests. However, to my surprise, the Type I error rate becomes underinflated for the RMANOVA as the effect size grows. The ANCOVA keeps a very accurate error rate, with an average Type I error rate of 5.0058% over 11 000 000 iterations, and with very little deviation (*SD* = 0.0233%). Over the same iterations, the RMANOVA's Type I error rate averages at 3.3367%, *SD* = 1.2639%. This result is very much influenced by the scenarios with a larger effect size. Instability in the Type I error rate is not something that a researcher wants hanging in the back of their thoughts when conducting analyses. This instability could only be justifiable if the RMANOVA were to bring forth significant parameter estimation advantages, which it does not.

When the correlation between Time 1 and Time 2 data approaches 1, the ANCOVA model gradually gains power. However, the main point of interest is the relative advantage of the ANCOVA versus the RMANOVA. As the correlation between the time points increases, or the effect size increases, the relative advantage of the ANCOVA over the RMANOVA decreases. The exception is with a near perfect correlation where a sudden spike in power occurs. A spike in power is also observed with the RMANOVA, but not to the same absolute level. I find it interesting to note that the relative gains within a method between the $r_{(t1,t2)} = 0.9$ and $r_{(t1,t2)} = 0.999$ cases favour the

ANCOVA as of $d = 0.2$ because of the plateau of total rejection of the null hypothesis. Making abstraction of this rapid shift when the correlation is too close to being perfect, the ANCOVA still maintains a healthy average of a 1.98:1 of successful rejection of the null hypothesis over the RMANOVA across all tests. The ANCOVA does not display any indication of instability in the Type I error rate, except when $r_{(t1,t2)} = 0.999$. I make this assessment based on a comparison of the ratio of Type I errors and successful rejections of the null hypothesis. Only when $r_{(t1,t2)} = 0.999$ is the ANCOVA's Type I error rate at a higher ratio than its advantage at successfully rejecting the null hypothesis. To be more precise, the Type I error rate ratio is 23 times higher than that of the RMANOVA, while the ratio of successful rejection of the null hypothesis is of 6.3:1 for the ANCOVA. The only way to capitalize on this would be to use an alpha of approximately 0.2 instead of 0.05 with a RMANOVA, but only when correlations are perfect. This is not something I would advise.

The last question of importance to assess in experiment 1 is the overlap of instances where the null hypothesis is rejected. The idea here is that if the ANCOVA does indeed reject the null at least twice as often when there is a known difference, do those rejections encompass all of the RMANOVA's rejections, or are they successfully rejecting instances where the data could have a particular and distinct pattern? Were this the case, it would raise its own set of very unique questions, but it is not. The overlap in rejection of the null hypothesis is nearly perfect. The peak discrepancy is of less than 1% when $r_{(t1,t2)} = 0.001$, but otherwise the cases that the RMANOVA rejects that the ANCOVA did not reject generally does not surpass 0.1% of the iterations.

When the sample size is increased and the same comparisons are made, we can see that the bottom limit of the Type I error rate for the RMANOVA is significantly closer to the 5% level at which it should be when compared to the scenarios where n = 20. When $r_{(t1,t2)}$ = 0.9, the scenario with n = 100 is approximately 50% closer to the correction Type I error rate than the scenario where n = 20 (3.7% and 2.1% respectively). However, as the $r_{(t1,t2)}$ tends towards 1, there is still a rapid drop to near 0 Type I errors. Furthermore, the difference between the two tests is much less visible outside of specific ranges of lower power, such as when d is smaller than 0.5.

Experiment 2: Determining the Sample Size Required to Obtain .80 Power

As hypothesized, the sample size requirements to reach a power of 0.80 for the ANCOVA are consistently smaller than for the RMANOVA. Requiring a sample size that is at least 1.5 times larger for the RMANOVA in order to reach the same level of power for a given scenario has major implications. Not only does this obviously increase costs and time requirements, it also means an experiment is more likely to fail due to other factors such as attrition, or procedural errors that lead to the loss of a part of the sample.

Experiment 3: Violating the Homogeneity of the Regression Slopes Assumption

As hypothesized, the ANCOVA keeps its advantage over the RMANOVA even when the homogeneity of the regression slopes assumption is violated. This is not an outright dismissal for the need to respect the assumption; again I re-iterate that the focus

here is on the relative strengths and weaknesses of the two methods being compared. The takeaway message here is that if the assumption of homogeneity of the regression slopes is important for the ANCOVA, it is even more important for the RMANOVA. As demonstrated through the examples with symmetrical and asymmetrical slopes, at the end of the day, the ANCOVA model uses the mean of the $r_{(t1,t2)}$ values for each of the two group levels. This means that regardless of the deviation, the mean of two exactly measured parameter estimates is still going to be a better overall estimate than fixing it to 1. Using an average of the slopes producing smaller residuals (being closer to the true value), than simply assuming that $r_{(t1,t2)} = 1$.

Experiment 4: Exploring the Effect of Large Systematic Baseline Differences

With the inflation of the ANCOVA's Type I error rate to as much as 40% incorrect rejections of the null hypothesis, the question of the overall power becomes moot in this scenario. If we can't trust the test to fail to reject the null hypothesis when there is no difference, then we can't trust subsequent rejections of the null hypothesis. As a matter of thoroughness, the patterns in rejections of the null hypothesis were still explored, and an interesting element crops up when the control group has a higher baseline score than the treatment group. Exceptionally, the RMANOVA outperforms the ANCOVA by approximately 20 percentage points, something not seen so far in the simulations. Not only that, but it occurs when there is a high effect size and a high baseline deviation. The high power situations are normally advantageous to the ANCOVA so far, but the baseline deviation is sufficient to counteract this. Combined

with the fact that the ANCOVA maintained its higher rate of rejections of the null hypothesis when the effect size was smaller, what we can conclude is that the ANCOVA is not simply more powerful than the RMANOVA when the effect size is large, but rather that the RMANOVA suffers when the effect size is small. If the opposite were the case, we would see a bottoming out of the scores in the low power high baseline deviation situation, which does not occur. One of the goals of the method used was to be able to draw a direct parallel between the different experiments. This allows me to add a further dimension to all other experiments when trying to understand and interpret the results; in low effect size situations, we are talking about a deflation of the RMANOVA's power, not some sort of unwarranted increase in the ANCOVA's.

Experiment 5: The Impact of Missing Data

Again, the ANCOVA does have more power to reject the null hypothesis. However, it was interesting to note that the ANCOVA's relative advantage over the RMANOVA is not as important when the control group ends up being larger than the treatment group. In estimating the control effect and the treatment effect, it would make sense that it is easier to estimate the parameters of the control group in this case, since the average does not change over time. Therefore, reduction in the total error component would be smaller when the precision is increased in the control group than when it is increased in the treatment group.

Difference between Interaction results and those of the Contingent Simple Effect

The contingent simple effect was chosen very specifically because I felt is had more of a practical application, but I am also very much aware that it can make the interpretation of the data in an absolute fashion somewhat more ambiguous and can be met with resistance. Since the goal was to explore the practical applications of the use of RMANOVA vs ANCOVA, there is a practical conclusion that can be drawn from having taken this approach. In cases of relatively low power (small sample size and small effect size), aside from the fact that researchers would still have drastically more power by using the ANCOVA, they run into another problem if they use the RMANOVA. They are face with an uncomfortably large probability that if the interaction term is statistically significant, they will remain unable to draw conclusions about where those differences lie. Of course it remains useful to know that the means changed differentially across groups, but if the results also indicate that there is no statistically significant difference between the groups at either time point, larger scale replication becomes almost necessary. The question “was this effect seen? because the treatment group improve more or was it the control group that improved more?” is not at all a trivial one. It is also important to note that all the code provided does allow researchers to use only the interaction terms should they desire to do so.

Merits of Additional Values to Test

One of the goals of this thesis was to factor in more possible values of a parameter being tested. This proved to be useful in showing how and when the differences between

the tests accelerated. One example is the increase rate at which the difference in the number of subjects needed to reach a power of .80 shrinks as the effect size grows. Another is the emergence of an ogive profile to the difference in the proportion of null hypotheses rejected by the ANCOVA and the RMANOVA, rather than miss-fitting the trend. With only two test values, the profile would be misinterpreted as being linear. With three or more values, where only one is on an extreme of the scale, profile would be misinterpreted as being quadratic. At least four well dispersed values, including both extremes, are necessary for the ogive profile to emerge.

Take Home

Another goal of this thesis was to provide a practical answer for researchers who simply want to know which test they can or should use. The general rules of thumb that can be discerned from the data gathered are the following. If there is a baseline imbalance, RMANOVA ***must*** be used. Research settings where there is an assumption of no baseline imbalance are those in which the grouping is non-random. The grouping variable must be the treatment; gender comparisons, cohort comparisons, clinically diagnosed vs control group, these are all situations where they may be no imbalance, but that state cannot be taken for granted. The use of ANCOVA in these kinds of situations greatly compromises the integrity of the conclusions even if there was no baseline imbalance, as it may affect replicability.

In a research setting with fully randomized group assignment, then the absence of a baseline imbalance can be assumed, and the following elements apply. If the projected

effect size is small, the ANCOVA will remain stable where the RMANOVA will lose power. If a research design was planned for an ANCOVA analysis, and constraints later lead the researcher to change to a difference score analysis, the sample size will have to be increased by a factor of 1.5 on average. If, on the other hand, a study was planned with a RMANOVA analysis in mind, and the use of ANCOVA becomes a possibility, fewer participants could be used. If there is homogeneity of the regression slopes issue, the ANCOVA should not be discounted on that basis alone. The assumption affects both methods, and does not lead to a sufficient change in the power profile of the two tests to shift the balance in favour of RMANOVA. Lastly, if there is data missing at random, or an uneven sample size, the ANCOVA should not be discounted on that basis alone. Although both tests are affected, the change in the power profile introduced by uneven sample size alone is not great enough to shift the balance in favour of the RMANOVA.

Limits

When $r_{(t1,t2)} = 1$

For all of the experiments, the correlation between Time 1 and Time 2 was limited to 0.999 because a perfect correlation caused issues in the computation of the models. The reason for this and the absence of a convergence between the two models when $r_{(t1,t2)} = 1$ was explored after the fact. An element that was perplexing as I reviewed the results is that as the correlation becomes close to perfect, the outcomes of the RMANOVA's interaction term and the ANCOVA's group difference term do not become identical as rapidly as expected. The relative advantage of the ANCOVA does shrink, indicating

some form of convergence, but without testing $r_{(t1,t2)} = 1$ exactly, it was impossible to know if they would eventually converge. Once the $r_{(t1,t2)}$ is set to 1, many of the values become difficult to interpret. All of a sudden, the incredibly stable Type I error rate of the ANCOVA is completely erratic. Some iterations also become unsolvable because a denominator could equal 0, and the overall Type I error rate of the ANCOVA goes up to 27.5%. With a very small effect size of $d = 0.1$ and $d = 0.2$, the ANCOVA rejects the null hypothesis more frequently than the RMANOVA, but that value remains relatively stable at approximately 50-51% and is utterly meaningless. On the other hand, the RMANOVA's interaction term's rate of rejection climbs linearly, but the Type I error rate is of 0. It is not feasible, with the method used, to generate data with both the required variance and a perfect correlation, thus making it impossible to reach a point where the results were identical for both methods.

Pooled variance in the data generation

It was suggested by Allison (1990) that multiple source of variability should be modeled into the data. In this case, all sources of variance were assumed to be combined within the variability introduced in the variance-covariance matrix. It was my view that once all the sources of variability are pooled together, they are not discernable in the analysis stage. However, it does mean that scenarios such as pre-test score based group assignment cannot be properly run with my program in its current state.

Next step

The current experiments are limited to a random assignment scenario. The next step is to explore the same parameters under a pre-test score group assignment paradigm.

Preexisting groups will remain out of the scope of the ANCOVA because of limits in the interpretability of the results. The other aspect that warrants further exploration is the effect of variance growth. The data generation model used in these experiments was unsuitable to explore the effect of variance growth, as changing the growth model pushed the correlation between Time 1 and Time 2 to extreme values.

Summary

It can be safely concluded that in fully randomized mixed models with no baseline differences, the ANCOVA is invariably better than a difference score in terms of statistical power. If there is a baseline difference, which should not occur in fully randomized research paradigms, then the ANCOVA fails in a very substantial fashion.

By using 1 000 000 iterations per scenario, the standard errors of the proportion of tests that fail or pass are inherently minuscule. Using the standard error equation for proportions,

$$SE_p = \sqrt{[p(1-p)/n]} \quad (6)$$

where p is the proportion of success, and n is the number of trials (iterations), we know that the standard error would never surpass 0.0005, or one twentieth of a percentage point at the peak error when $p = 0.5$. From this we can conclude that the success rates provided in the results can be taken nearly as is, for a +- 0.001 (2 *SE*s) specification would not change in any way the conclusions that were drawn. This also means that we can affirm that when the Type I error rate of the RMANOVA is lower than 5%, there is, in fact, a problem. Kisbu-Sakarya et al. (2012) reported Type I error rates of 0.04 to 0.07 as being

no different to 0.05, but we can clearly see that even when the high number of simulations stabilizes the values, they do not recover to an acceptable value in the case of the RMANOVA.

Allowing supplementary parameters to vary is normally seen as something that can increase the precision of a parameter estimation function, but at the cost of a greater sample size requirement. This is why fixing parameters is of great use in complex models such as structural equation models. Sample sizes need to be greatly increased, and researchers go to great lengths to be able to gain better estimates in those cases. Here we have a low hanging fruit. At the negative cost of lower sample size requirements, greater precision can be achieved in the parameter estimates of mixed model research paradigms by using an ANCOVA, or, in other words, allowing B_1 to vary.

There are certainly many elements that can reduce the power of the ANCOVA; assumptions to respect, etc. However, by systematically comparing the ANCOVA to the RMANOVA, we can see that those concerns, valid as they may be, simply do not affect the ANCOVA enough to render it less accurate that the RMANOVA. If the situation is such that the assumption violations are so severe that the ANCOVA is not usable, then neither is the RMANOVA. As the situation becomes less statistically demanding, the ANCOVA does lose some of its relative advantage over the RMANOVA, but nowhere nearly enough to warrant the use of the RMANOVA over the ANCOVA. Lastly, it is with the most demanding situations (smaller effect sizes, weaker correlations), that the ANCOVA can easily halve the number of participants required to obtain the same outcome.

Glossary

B: Unstandardized regression coefficient.

β: Standardized regression coefficient.

Change Score: see Difference Score.

Contingent Power of the Simple Effect: considering a rejection of the null hypothesis of any of an interaction's simple effects, contingent on that interaction term having been statistically significant in the first place.

Δ: See Difference Score. Shorthand used in equations and mathematical notations.

ε: Measurement error.

Estimated Marginal Mean: Estimated mean after adjusting for the covariate.

Deficiency Score: See Difference Score.

Difference Score: A score that is obtained by subtracting the first score from the second score. Same a change score, deficiency score, Delta, Δ.

***G*:** See Group. Shorthand used in equations and mathematical notations.

Group / Grouping Variable: Variable which defines assignment to the treatment or control group in most cases (when there is no control, it will be specified). Used interchangeably with treatment level.

Growth – Fan Spread: Fan spread growth refers to the case in which the difference between the Time 2 and Time 1 scores is positively related with the Time 1 score.

Growth – Mastery: Mastery growth refers to the case in which the difference between the Time 2 and Time 1 scores is negatively related with the Time 1 score

Growth – Parallel: Parallel growth refers to the case in which the difference between the Time 2 and Time 1 scores is inconsequentially related with the Time 1 score

Iterations: The number of times specific parameter settings for a given scenario of a simulation is tested.

Monte Carlo: This approach consists of repeatedly creating random samples, and analyzing these samples.

Residualized (Change Score): Method that was a step towards ANCOVA. The regression weight of the pooled data is used rather than the averaged regression weight of each group, causing poor estimations of the parameter and an important bias and loss of power.

***T_1*:** See Time 1. Shorthand used in equations and mathematical notations.

***T_2*:** See Time 2. Shorthand used in equations and mathematical notations.

Scenario: Refers to a given set of conditions that will be explored within a simulation. A situation under which a specific parameter or group of parameters will be varied throughout the iterations.

Simulation: Refers to the method of generating data for the sake of testing a model with given parameters. Includes, but is not limited to Monte Carlo simulations.

Time 1: In the context of this work, it is the used interchangeably with pre-test, pretest, baseline, and T_1.

Time 2: In the context of this work, it is used interchangeably with post-test, post-test, outcome, and T_2.

Treatment Level: See Group.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93. http://doi.org/10.2307/271083

Cribbie, R. a., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*, *60*(6), 893–907. http://doi.org/10.1007/s13398-014-0173-7.2

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": or should we? *Psychological Bulletin*, *74*(1), 68–80. http://doi.org/10.1037/h0029382

DigitalOcean. (2016). DigitalOcean. Retrieved from https://www.digitalocean.com/

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 265–287. http://doi.org/10.1177/109442810143005

Forbes, A. B., & Carlin, J. B. (2005). "Residual change" analysis is not equivalent to analysis of covariance. *Journal of Clinical Epidemiology*, *58*(5), 540–541. http://doi.org/10.1016/j.jclinepi.2004.12.002

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance*. http://doi.org/10.1016/0030-5073(81)90033-7

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2012). A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*, *73*(1), 47–62. http://doi.org/10.1177/0013164412450574

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. (B. Gordon, R. T. Hercher, & L. Stone, Eds.) (Fifth Edit). New York: McGraw-Hill Irwin.

Lord, F. M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.), *Proglems in Measuring Change* (First Edit). Madison: University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304–305. http://doi.org/10.1037/h0025105

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods*. http://doi.org/10.1037/1082-989X.3.3.309

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48.

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*(1), 3–28. http://doi.org/10.1177/0193841X0102500101

Overall, J. E., & Doyle, S. R. (1994a). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials*, *123*(15), 100–123.

Overall, J. E., & Doyle, S. R. (1994b). Implications of chance baseline differences in repeated measurement designs. *Journal of Biopharmaceutical Statistics*, *4*(2), 199–216.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*(1), 85–86. http://doi.org/10.1037/h0076158

Owen, S. V, & Froman, R. D. (1998). Focus on qualitative methods Uses and abuses of the analysis of covariance. *Research in Nursing & Health*, *21*, 557–562.

Petscher, Y., & Schatschneider, C. (2011). A simulation study on the performance of the simple difference and covariance-adjusted scores in randomized experimental designs. *Journal of Educational Measurement*, *48*(1), 31–43. http://doi.org/10.1111/j.1745-3984.2010.00129.x

Porter, L. W. (1962). Job attitudes in management: Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology*, *46*(6), 375–384.

R Core Team. (2016). R. Retrieved from https://www.r-project.org/

RStudio. (2016). RStudio Server. Retrieved from https://www.rstudio.com/products/rstudio/download-server/

Thomas, D. R., & Zumbo, B. D. (2011). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. http://doi.org/10.1177/0013164411409929

Van Breukelen, G. J. . (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*, 920–925. http://doi.org/10.1016/j.jclinepi.2006.02.007

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again.

Psychological Bulletin, *109*(1), 147–151. http://doi.org/10.1037/0033-2909.109.1.147

Wall, T. D., & Payne, R. (1973). Are deficiency scores deficient? *Journal of Applied Psychology*, *58*(3), 322–326. http://doi.org/10.1037/h0036227

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*. http://doi.org/10.1080/00031305.2016.1154108

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*(1), 59–69. http://doi.org/10.1177/014662169602000106

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles in Experimental Design*. McGraw-Hill. Retrieved from https://books.google.ca/books/about/Statistical_Principles_in_Experimental_D.html?id=OqppAAAAMAAJ&pgis=1

Wright, D. B. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology*, *76*(3), 663–675. http://doi.org/10.1348/000709905X52210

Zimmerman, D. W. (2009). The reliability of difference scores in populations and samples. *Journal of Educational Measurement*, *46*(1), 19–42. http://doi.org/10.1111/j.1745-3984.2009.01067.x

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*(2), 149–154. http://doi.org/10.1111/j.1745-3984.1982.tb00124.x

Zimmerman, D. W., Williams, R. H., Zumbo, B. D., Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Significance tests based on differences. *Applied Psychological Measurement*, *17*(1), 1–9. http://doi.org/10.1177/014662169301700101

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. *Advances in Social Science Methodology*, *5*, 269–304.

Appendix 1: Main R Code for the Simulations

```
#comments were kept in code to help with clarity. All different steps of the scenarios
were written in instead of using a source parameter in order to document each step and
facilitate the management of the 20 instances of R Studio Server, troubleshooting, and
documentation of progress over extended periods of time. In order to run a given
scenario, the appropriate conditions would be uncommented, and the program launched.

library(compiler); enableJIT(3)
start<-Sys.time(); start #timestamp to know the time needed
library(MASS); set.seed(NULL) #main seed is 15654, verify in case changed for
troubleshooting purposes
i<-0; j<-0; dincreases<-12; loops<-1000000 #loop parameters and repetitions; loops is
number of time for same d and j is it's counter; defaults are 12 and 25000, change to
test stuff

#Exp 1 (use baseline means from exp 5 for the means); is new exp 1
#correlation tends towards 1, should even out deltaT and ANCOVA
#set fixed parameters; rmat is the covariance matrix, using 15 squared (for the SD of
15), multiplied by r (the correlation coefficient)
#rmat <- matrix(c(15*15, 15*15*0.001, 15*15*0.001, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.001, 15*15*0.001, 15*15),2,2); exper<-"3b_01"
#rmat <- matrix(c(15*15, 15*15*0.1, 15*15*0.1, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.1, 15*15*0.1, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.2, 15*15*0.2, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.2, 15*15*0.2, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.3, 15*15*0.3, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.3, 15*15*0.3, 15*15),2,2)
```

```
#rmat <- matrix(c(15*15, 15*15*0.4, 15*15*0.4, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.4, 15*15*0.4, 15*15),2,2)
rmat <- matrix(c(15*15, 15*15*0.5, 15*15*0.5, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.5, 15*15*0.5, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.6, 15*15*0.6, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.6, 15*15*0.6, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.7, 15*15*0.7, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.7, 15*15*0.7, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.8, 15*15*0.8, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.8, 15*15*0.8, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.9, 15*15*0.9, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.9, 15*15*0.9, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.999, 15*15*0.999, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.999, 15*15*0.999, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15, 15*15, 15*15),2,2); rmat2 <- matrix(c(15*15, 15*15,
15*15, 15*15),2,2) #for testing r=1
#requires the additional removal of NANs from the dataset with following code that has to
be run after the fact
  #Ap05b<-Ap; Apb<-Ap
  #Ap05b[is.na(Ap05b)]<-0; #value can be changed for sig. or not.
  #Ap05b[Apb<0.05]<-1;        Ap05b[Apb>0.05]<-0
  #colMeans (Ap05b)

#Epx 3, homogeneity of regression slopes violation (use baseline means from exp 5)
#rmat <- matrix(c(15*15, 15*15*(-0.1), 15*15*(-0.1), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.1, 15*15*0.1, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.3), 15*15*(-0.3), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.3, 15*15*0.3, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.5), 15*15*(-0.5), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.5, 15*15*0.5, 15*15),2,2)
```

```
#rmat <- matrix(c(15*15, 15*15*(-0.7), 15*15*(-0.7), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.7, 15*15*0.7, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.9), 15*15*(-0.9), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.9, 15*15*0.9, 15*15),2,2) #rmat <- matrix(c(15*15, 15*15*0.3,
15*15*0.3, 15*15),2,2); rmat2 <- matrix(c(15*15, 15*15*0.5, 15*15*0.5, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.5, 15*15*0.5, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.3, 15*15*0.3, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.5, 15*15*0.5, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.7, 15*15*0.7, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*0.1, 15*15*0.1, 15*15),2,2); rmat2 <- matrix(c(15*15,
15*15*0.7, 15*15*0.7, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.3), 15*15*(-0.3), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.9, 15*15*0.9, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.9), 15*15*(-0.9), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.3, 15*15*0.3, 15*15),2,2) #rmat <- matrix(c(15*15, 15*15*(0.9),
15*15*(0.9), 15*15),2,2); rmat2 <- matrix(c(15*15, 15*15*-0.3, 15*15*-0.3, 15*15),2,2)
#rmat <- matrix(c(15*15, 15*15*(-0.6), 15*15*(-0.6), 15*15),2,2); rmat2 <-
matrix(c(15*15, 15*15*0.6, 15*15*0.6, 15*15),2,2)


#means2 is means2s for start; duplicate variable prevents use of duplicate typed values,
reducing the chances of human error as variables self-propagate
# each = number of subjects per group
#EXP 4 as group 2 t2 will keep changing at it's normal intervals, (use exp 1 base
covariance matrix of 0.5)
means <- c(100,100); means2s <- c(100,100) #baseline means; exp1 baseline + this baseline
also equals experiments 4a_00 and 4b_00

#4A- group 2 is higher at baseline
#means2s <- c(101.5,101.5); means <- c(100,100);
#means2s <- c(103,103);     means <- c(100,100);
```

```
#means2s <- c(104.5,104.5); means <- c(100,100);
#means2s <- c(106,106);     means <- c(100,100);
#means2s <- c(107.5,107.5); means <- c(100,100);
#means2s <- c(109,109);     means <- c(100,100);
#means2s <- c(110.5,110.5); means <- c(100,100);
#means2s <- c(112,112);     means <- c(100,100);
#means2s <- c(113.5,113.5); means <- c(100,100);
#means2s <- c(115,115);     means <- c(100,100);


#4B- group 1 is higher at baseline
#means <- c(101.5,101.5); means2s <- c(100,100);
#means <- c(103,103);     means2s <- c(100,100);
#means <- c(104.5,104.5); means2s <- c(100,100);
#means <- c(106,106);     means2s <- c(100,100);
#means <- c(107.5,107.5); means2s <- c(100,100);
#means <- c(109,109);     means2s <- c(100,100);
#means <- c(110.5,110.5); means2s <- c(100,100);
#means <- c(112,112);     means2s <- c(100,100);
#means <- c(113.5,113.5); means2s <- c(100,100);
#means <- c(115,115);     means2s <- c(100,100);


#exp5 for missingness, which is actually a sample size imbalance experiment will rely
entirely on modifying the subgr variable
#use exp1  matrix of 0.5 and exp4 baseline
#number of subjects per group for group 1; #number of subjects per group for group 2
subgr<-20; subgr2<-20 #baseline
#subgr<-20; subgr2<-21
#subgr<-20; subgr2<-22
#subgr<-20; subgr2<-23
```

```
#subgr<-20; subgr2<-24
#subgr<-20; subgr2<-25
#subgr<-20; subgr2<-26
#subgr<-20; subgr2<-27
#subgr<-20; subgr2<-28
#subgr<-20; subgr2<-29
#subgr<-20; subgr2<-30
#subgr<-21; subgr2<-20
#subgr<-22; subgr2<-20
#subgr<-23; subgr2<-20
#subgr<-24; subgr2<-20
#subgr<-25; subgr2<-20
#subgr<-26; subgr2<-20
#subgr<-27; subgr2<-20
#subgr<-28; subgr2<-20
#subgr<-29; subgr2<-20
#subgr<-30; subgr2<-20


#rep(1:number of subjects total)
subgrtot<-subgr+subgr2 #number of data points per group since there are 2 time points;
used to size and navigate matrices
subgrpl1<-subgr+1 #group size + 1; used to keep the reading in sequence in matrices
subgr2pl1<-subgr2+1 #group 2 size + 1; used to keep the reading in sequence in matrices
rmdf<-subgr*2-2 #degrees of freedom for RM
adf<-subgr*2-3 #DFs for ancova
Subject<- rep(1:subgrtot,times=1)
G1<-rep(c(1),each=subgr,times=1) #creates group 1 tags
G2<-rep(c(2),each=subgr2,times=1) #creates group 2 tags
G<-c(G1,G2) #combines them into a single group tag vector
```

```
#declare some variables for the loop
mat=NULL; RMmat=NULL; mat1=NULL; mat2=NULL; acp=NULL; means2=means2s;
Ap<-matrix(data=NA,nrow=loops,ncol=(dincreases+1))
RMp<-matrix(data=NA,nrow=loops,ncol=(dincreases+1))
RMSp<-matrix(data=NA,nrow=loops,ncol=(dincreases+1)) #used for the collection of p values
for the time 2 simple effect, no suffix because it was originally the only one
RMSp2<-matrix(data=NA,nrow=loops,ncol=(dincreases+1)) #use for the collection of p values
for time 1 simple effect
growthcor<-matrix(data=NA,nrow=loops,ncol=(dincreases+1)) #collects t2-t1,t1 correlation
values (averaged across both groups)
timecor<-matrix(data=NA,nrow=loops,ncol=(dincreases+1))   #collects t1,t2 correlation
values (averaged across both groups)
g1d<-matrix(data=NA,nrow=loops,ncol=(dincreases+1))       #collects actual average
cohen's d for group 1
g2d<-matrix(data=NA,nrow=loops,ncol=(dincreases+1))       #collects actual average
cohen's d for group 2

while (j<loops) {
  i<-0
  while (i<dincreases+1){
    rpos=j+1 #loop position in table
    cpos=i+1 #d position in table
    #clear variables to avoid bugs
    mat=NULL; mat2=NULL; mat=NULL; T1= NULL; T2= NULL; dataset=NULL; SCE=NULL; RMT2=NULL;
SCT=NULL
    means2[2]<-means2s[2]+(i*1.5) #set dynamic matrix parameters for increasing means

    #generate data using mvnorm, rmat is sigma, the matrix specifying the covairance
matrix of the variables
    # mvnorm(number of jusbects per group,...) twice
```

```
    mat1 <- mvrnorm(subgr, means, rmat, TRUE); mat2 <- mvrnorm(subgr2, means2, rmat2,
TRUE)
     #mat 1 is group 1 at T1 and T2; while mat2 is group 2;
      #they are independently generated, but within the group are correlated using the
covariance matrix rmat
    #combine data for analysis
    # dataset matrix is of length of all subjects
    mat<-rbind(mat,mat1,mat2); T1<-mat[,1]; T2<-mat[,2]; dataset<-
matrix(c(G,T1,T2,Subject),(subgrtot),4)

    growthcor[rpos,cpos]<- (cor((mat1[,2]-mat1[,1]),mat1[,1])+cor((mat2[,2]-
mat2[,1]),mat2[,1]))/2
    timecor[rpos,cpos]<- (cor(mat1[,1],mat1[,2])+cor(mat2[,1],mat2[,2]))/2

    g1d[rpos,cpos]<- mean(((mat1[,2]-mat1[,1])/15))
    g2d[rpos,cpos]<- mean(((mat2[,2]-mat2[,1])/15))

    #ijk is ith observation of jth group of kth time, while i is mean of subject
    # ranges used correspond to number of subjects per group, at the end of the function
)),total data points so subjects by times,1)
    jkmeans<-
matrix(c(mean(dataset[1:subgr,2]),mean(dataset[subgrpl1:subgrtot,2]),mean(dataset[1:subgr
,3]),mean(dataset[subgrpl1:subgrtot,3])),2,2)
    grmean=mean(c(jkmeans[1,],jkmeans[2,])); meank1=mean(jkmeans[,1])
;meank2=mean(jkmeans[,2]); meanj1=mean(jkmeans[1,]); meanj2=mean(jkmeans[2,])
    indiv<-matrix(c(dataset[,2],dataset[,3]),(subgrtot*2),1)
    #multiplier at begining is number of subjects per group; SS of the G T interaction
    SSGT<-subgr*((jkmeans[1,1]-meanj1-meank1+grmean)^2+(jkmeans[1,2]-meanj1-
meank2+grmean)^2)+
          subgr2*((jkmeans[2,1]-meanj2-meank1+grmean)^2+(jkmeans[2,2]-meanj2-
meank2+grmean)^2)
```

```
    #                           ij  k            .jk                          ij.
...
    SSEGT<-sum(((matrix(c(dataset[1:subgr,2]),subgr,1))-jkmeans[1,1]-
(rowSums((matrix(c(dataset[1:subgr,2],dataset[1:subgr,3]),subgr,2)))/2)+meanj1)^2+
                  ((matrix(c(dataset[1:subgr,3]),(subgr),1))-jkmeans[1,2]-
(rowSums((matrix(c(dataset[1:subgr,2],dataset[1:subgr,3]),(subgr),2)))/2)+meanj1)^2)+
            sum(((matrix(c(dataset[subgrpl1:subgrtot,2]),(subgr2),1))-jkmeans[2,1]-
(rowSums((matrix(c(dataset[subgrpl1:subgrtot,2],dataset[subgrpl1:subgrtot,3]),(subgr2),2)
))/2)+meanj1)^2+
                  ((matrix(c(dataset[subgrpl1:subgrtot,3]),(subgr2),1))-jkmeans[2,2]-
(rowSums((matrix(c(dataset[subgrpl1:subgrtot,2],dataset[subgrpl1:subgrtot,3]),(subgr2),2)
))/2)+meanj1)^2)

    RMp[rpos,cpos]<-(pf((SSGT/(SSEGT/rmdf)),1,rmdf,lower.tail=FALSE)) #MSG:T is 1 df, so
same as SS
    #other parts of the RMANOVA not currently in use, so not computed, skip to simple
effects; RMSp is time 1 simple effect and RMSp2 is Time 2 simple effect
    RMSp[rpos,cpos]<-(pf(((subgr*((jkmeans[1,2]-meank2)^2+(jkmeans[2,2]-
meank2)^2))/(SSEGT/rmdf)),1 ,rmdf,lower.tail=FALSE))
    RMSp2[rpos,cpos]<-(pf(((subgr*((jkmeans[1,1]-meank1)^2+(jkmeans[2,1]-
meank1)^2))/(SSEGT/rmdf)),1 ,rmdf,lower.tail=FALSE))

    #for the manual ANCOVA, I will need SS for within each time within each group, as
well as the SUM of the cross product of T1T2 within each group
    SSGT2<-sum(((matrix(c(dataset[1:subgrtot,3]),(subgrtot),1))-meank2)^2)
    SSGT1<-sum(((matrix(c(dataset[1:subgrtot,2]),(subgrtot),1))-meank1)^2)
    SScodevT<-
(sum((matrix(c(dataset[1:subgrtot,2]),(subgrtot),1))*(matrix(c(dataset[1:subgrtot,3]),(su
bgrtot),1))))-
((sum(matrix(c(dataset[1:subgrtot,3]),(subgrtot),1)))*(sum(matrix(c(dataset[1:subgrtot,2]
),(subgrtot),1)))/subgrtot)
```

```
    SSGT2E<-sum((matrix(c(dataset[1:subgr,3]),(subgr),1)-
jkmeans[1,2])^2)+sum((matrix(c(dataset[subgrpl1:subgrtot,3]),(subgr2),1)-jkmeans[2,2])^2)
    SSGT1E<-sum((matrix(c(dataset[1:subgr,2]),(subgr),1)-
jkmeans[1,1])^2)+sum((matrix(c(dataset[subgrpl1:subgrtot,2]),(subgr2),1)-jkmeans[2,1])^2)
    SScodevE<-
sum(matrix(c(dataset[subgrpl1:subgrtot,2]))*matrix(c(dataset[subgrpl1:subgrtot,3])))-
(sum(matrix(c(dataset[subgrpl1:subgrtot,2])))*sum(matrix(c(dataset[subgrpl1:subgrtot,3]))
)/subgr2)+
      sum(matrix(c(dataset[1:subgr,2]))*matrix(c(dataset[1:subgr,3])))-
(sum(matrix(c(dataset[1:subgr,2])))*sum(matrix(c(dataset[1:subgr,3])))/subgr)
    #adjSSTR<-(SSGT2-(SScodevT^2/SSGT1))-(SSGT2E-(SScodevE^2/SSGT1E))
    #adjSST=SSGT2-(SScodevT^2/SSGT1)
    #adjSSE=SSGT2E-(SScodevE^2/SSGT1E)
    Ap[rpos,cpos]<-(pf((((SSGT2-(SScodevT^2/SSGT1))-(SSGT2E-
(SScodevE^2/SSGT1E)))/((SSGT2E-(SScodevE^2/SSGT1E))/adf)),1 ,adf,lower.tail=FALSE))

    # move forward
    i<-i+1;
  }
  j<-j+1; #print(i)
}
#flag of repeated measure interaction being statistically significant
RMp05<-RMp;     RMp05[RMp<0.05]<-1;      RMp05[RMp>0.05]<-0
#simple effect conditional on interaction when "#" are removed from in front of the 4th
column
#flag of repeated measure T2 simple effect being statistically significant
RMSp05<-RMSp;   RMSp05[RMSp<0.05]<-1;    RMSp05[RMSp>0.05]<-0;      RMSp05[RMp>0.05]<-0
#flag of repeated measure T1 simple effect being statistically significant
RMSp205<-RMSp2; RMSp205[RMSp2<0.05]<-1;  RMSp205[RMSp2>0.05]<-0;    RMSp205[RMp>0.05]<-0
#flag of RM T1 OR T2 being statistically significant contingent on RM
```

```
RMSboth05<-RMSp05; RMSboth05[RMSp205=1]<-1; RMSboth05[RMp>0.05]<-0
#flag of ANCOVA group difference being staitistically significant
Ap05<-Ap;       Ap05[Ap<0.05]<-1;        Ap05[Ap>0.05]<-0


#conditional: RMSp is not sig but Ap is
CAp05<-matrix(0,nrow=loops,ncol=(dincreases+1));
CAp05[(RMSp05==0)&(Ap05==1)]<-1

#conditional: Ap is not sig, but RMSp is
CRMSp05<-matrix(0,nrow=loops,ncol=(dincreases+1)); CRMSp05[(RMSboth05==1)&(Ap05==0)]<-1 #
two time point conditional significance

#collects count of significance as a % of significant tests for RM interaction, T2 simple
effect, and ANCOVA group significance
p05<-matrix((colMeans(RMp05)),nrow=1, ncol=(dincreases+1), byrow=TRUE);
p05<-rbind(p05,(colMeans(RMSboth05)),(colMeans(Ap05)),colMeans(CAp05),colMeans(CRMSp05))
#two time point conditional significance
pdifwin<-matrix((p05[3,]-p05[2,]),nrow=1,ncol=(dincreases+1));
p05<-
rbind(p05,pdifwin,(colMeans(growthcor)),(colMeans(timecor)),colMeans(g1d),colMeans(g2d))

#name rows and columns
rownames(p05)<-c("RMANOVA_0.05","RM-
T2_simple_effect","ANCOVA_0.05","Ap_sig_but_not_RMSp","RMSp_sig_but_not_Ap","ANCOVA_minus
_RM", "mean_growth_cor", "mean_t1t2_cor","g1_d","g2_d");
colnames(p05)<-seq(from=0, to=(dincreases*0.1), by = 0.1)

#table with Ap and RMS respective wins
win<-matrix(p05[4,],nrow=1,ncol=(dincreases+1)); win<-rbind(win,p05[5,])
rownames(win)<-c("A_sig_but_not_RMs_p05","RMS_sig_but_not_A_p05")
```

```
#table with Ap and RMS respective wins; reorganized
win2<-matrix(p05[4,],nrow=1,ncol=(dincreases+1));
win2<-rbind(win2,p05[5,])
rownames(win2)<-c("A_sig_but_not_RMs_p05","RMS_sig_but_not_A_p05")


finish<-Sys.time() #timestamp to know the time needed
#the simulation required a
(finish-start)
#for a total of
(loops*(dincreases+1))
#loops#p1;
p05
```

Appendix 2: R Code for Experiment 2

```
library(compiler); enableJIT(3)
start<-Sys.time(); start #timestamp to know the time needed
library(MASS); set.seed(NULL) #main seed is 15654 for troubleshooting purposes
n <- 0; i<-0; j<-0; nincreases<-19; loops<-1000000 #loop parameters and repetitions;
loops is number of time for same d; defaults are 19 and 25000, change to test stuff

#set fixed parameters; rmat is the covariance matrix, using 15 squared (for the SD of
15), multiplied by r (the correlation coefficient)
rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2)
means <- c(100,100); means2 <- c(100,100);   exper<-"2_01" # small effect size = 104.5
med = 109, large = 113.5
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,101.5)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,103)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,104.5)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,106)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,107.5)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,109)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,110.5)
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,112)
```

```
#rmat <- matrix(c(15*15, 15*15*0.50, 15*15*0.50, 15*15),2,2); means <- c(100,100); means2
<- c(100,113.5)

#declare some variables for the loop
mat=NULL; RMmat=NULL; ANCOVAp=NULL; RMANOVAp=NULL; RMANOVAMp=NULL; RMANOVAGp=NULL;
RMST2p=NULL; mat1=NULL; mat2=NULL; ANOVAp=NULL; acp=NULL
Ap<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))
RMp<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))
RMMp<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))
RMGp<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))
RMSp<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))
RMSp1<-matrix(data=NA,nrow=loops,ncol=(nincreases+1))

while (i<nincreases+1) {
  j<-0
  n<- n + 10
  G<- rep(c(1,2), each=n, times=1); Subject<- rep(1:(n*2),times=1) #
  while (j<loops){
    rpos=j+1 #loop position in table
    cpos=i+1 #n size position in table
    mat=NULL; mat2=NULL; mat=NULL; T1= NULL; T2= NULL; dataset=NULL; SCE=NULL; RMT2=NULL;
SCT=NULL ; deltaT=NULL

    #generate data using mvnorm, rmat is sigma, the matrix specifying the covariance
matrix of the variables
    mat1 <- mvrnorm(n, means, rmat, TRUE); mat2 <- mvrnorm(n, means2, rmat, TRUE) #mat 1
is group 1 at T1 and T2; while mat2 is group 2;
    #they are independently generated, but within the group are correlated using the
covariance matrix rmat
    #combine data for analysis
```

```
    mat<-rbind(mat,mat1,mat2); T1<-mat[,1]; T2<-mat[,2]; deltaT<-(mat[,2]-mat[,1]);
dataset<-matrix(c(G,T1,T2,Subject),(n*2),4)

    #ijk is ith observation of jth group of kth time, while i is mean of subject
    # ranges used correspond to number of subjects per group, at the end of the
function)),total data points so subjects by times,1)
    jkmeans<-
matrix(c(mean(dataset[1:n,2]),mean(dataset[(n+1):(n*2),2]),mean(dataset[1:n,3]),mean(data
set[(n+1):(n*2),3])),2,2)
    grmean=mean(c(jkmeans[1,],jkmeans[2,])); meank1=mean(jkmeans[,1])
;meank2=mean(jkmeans[,2]); meanj1=mean(jkmeans[1,]); meanj2=mean(jkmeans[2,])
    indiv<-matrix(c(dataset[,2],dataset[,3]),(n*4),1)
    #multiplier at begining is number of subjects per group
    SSGT<-n*((jkmeans[1,1]-meanj1-meank1+grmean)^2+(jkmeans[2,1]-meanj2-meank1+grmean)^2+
             (jkmeans[1,2]-meanj1-meank2+grmean)^2+(jkmeans[2,2]-meanj2-meank2+grmean)^2)
    #                          ij  k           .jk                         ij.
...
    SSETGT<-sum(((matrix(c(dataset[1:n,2]),(n),1))-jkmeans[1,1]-
(rowSums((matrix(c(dataset[1:n,2],dataset[1:n,3]),(n),2)))/2)+meanj1)^2+
                  ((matrix(c(dataset[1:n,3]),(n),1))-jkmeans[1,2]-
(rowSums((matrix(c(dataset[1:n,2],dataset[1:n,3]),(n),2)))/2)+meanj1)^2+
                  ((matrix(c(dataset[(n+1):(n*2),2]),(n),1))-jkmeans[2,1]-
(rowSums((matrix(c(dataset[(n+1):(n*2),2],dataset[(n+1):(n*2),3]),(n),2)))/2)+meanj1)^2+
                  ((matrix(c(dataset[(n+1):(n*2),3]),(n),1))-jkmeans[2,2]-
(rowSums((matrix(c(dataset[(n+1):(n*2),2],dataset[(n+1):(n*2),3]),(n),2)))/2)+meanj1)^2)
    RMp[rpos,cpos]<-(pf((SSGT/(SSETGT/(n*2-2))),1,(n*2-2),lower.tail=FALSE)) #MSG:T is 1
df, so same as SS
    RMSp[rpos,cpos]<-(pf(((n*((jkmeans[1,2]-meank2)^2+(jkmeans[2,2]-
meank2)^2))/(SSETGT/(n*2-2))),1,(n*2-2),lower.tail=FALSE))
    RMSp1[rpos,cpos]<-(pf(((n*((jkmeans[1,1]-meank1)^2+(jkmeans[2,1]-
meank1)^2))/(SSETGT/(n*2-2))),1,(n*2-2),lower.tail=FALSE))
```

```
    #for the manual ancova, I will need SS for within each time within each group, as
well as the SUM of the cross product of T1T2 within each group
    SSGT2<-sum(((matrix(c(dataset[1:(2*n),3]),((2*n)),1))-meank2)^2)
    SSGT1<-sum(((matrix(c(dataset[1:(2*n),2]),((2*n)),1))-meank1)^2)
    SScodevT<-
(sum((matrix(c(dataset[1:(2*n),2]),((2*n)),1))*(matrix(c(dataset[1:(2*n),3]),((2*n)),1)))
)-
((sum(matrix(c(dataset[1:(2*n),3]),((2*n)),1)))*(sum(matrix(c(dataset[1:(2*n),2]),((2*n))
,1)))/(2*n))

    SSGT2E<-sum((matrix(c(dataset[1:n,3]),(n),1)-
jkmeans[1,2])^2+(matrix(c(dataset[(n+1):(2*n),3]),(n),1)-jkmeans[2,2])^2)
    SSGT1E<-sum((matrix(c(dataset[1:n,2]),(n),1)-
jkmeans[1,1])^2+(matrix(c(dataset[(n+1):(2*n),2]),(n),1)-jkmeans[2,1])^2)
    SScodevE<-sum(matrix(c(dataset[(n+1):(2*n),2]))*matrix(c(dataset[(n+1):(2*n),3])))-
(sum(matrix(c(dataset[(n+1):(2*n),2])))*sum(matrix(c(dataset[(n+1):(2*n),3])))/n)+
      sum(matrix(c(dataset[1:n,2]))*matrix(c(dataset[1:n,3])))-
(sum(matrix(c(dataset[1:n,2])))*sum(matrix(c(dataset[1:n,3])))/n)
    #adjSSTR<-(SSGT2-(SScodevT^2/SSGT1))-(SSGT2E-(SScodevE^2/SSGT1E))
    #adjSST=SSGT2-(SScodevT^2/SSGT1)
    #adjSSE=SSGT2E-(SScodevE^2/SSGT1E)
    Ap[rpos,cpos]<-(pf((((SSGT2-(SScodevT^2/SSGT1))-(SSGT2E-
(SScodevE^2/SSGT1E)))/((SSGT2E-(SScodevE^2/SSGT1E))/(2*n-3))),1 ,(2*n-
3),lower.tail=FALSE))

    # move forward
    j<-j+1;
  }
  i<-i+1; #print(i)
  G=NULL; Subject=NULL;
```

```
}

#creates ordered lists for significant p counts; simple effect can also be conditional on
interaction justifying looking into it
RMp05<-RMp;  RMp05[RMp<0.05]<-1;    RMp05[RMp>0.05]<-0

#testing for any Type I error in the G and M and combined
RMGp05<-RMGp;  RMGp05[RMGp<0.05]<-1;    RMGp05[RMGp>0.05]<-0
#RMMp1<-RMMp;   RMMp1[RMMp<0.1]<-1;      RMMp1[RMMp>0.1]<-0
RMMp05<-RMMp;  RMMp05[RMMp<0.05]<-1;    RMMp05[RMMp>0.05]<-0
RMAp05<-RMMp05+RMGp05+RMp05;        RMAp05[RMAp05!=0]<-1;    RMAp05[RMAp=0]<-0 #RMAP no
longer used, leftover from older code. to remove in cleanup.

#simple effect conditional on interaction when "#" are removed from in front of the 4th
column
#flag of repeated measure T2 simple effect being statistically significant
RMSp05<-RMSp;  RMSp05[RMSp<0.05]<-1;    RMSp05[RMSp>0.05]<-0;    RMSp05[RMp>0.05]<-0
#flag of repeated measure T1 simple effect being statistically significant
RMSp105<-RMSp1; RMSp105[RMSp1<0.05]<-1;  RMSp105[RMSp1>0.05]<-0;    RMSp105[RMp>0.05]<-0
#flag of RM T1 OR T2 being statistically significant contingent on RM
RMSboth05<-RMSp05; RMSboth05[RMSp105=1]<-1; RMSboth05[RMp>0.05]<-0
#flag of ANCOVA group difference being staitistically significant
Ap05<-Ap;    Ap05[Ap<0.05]<-1;      Ap05[Ap>0.05]<-0

#conditional: RMSp is not sig but Ap is
CAp05<-matrix(0,nrow=loops,ncol=(nincreases+1));
CAp05[(RMSp05==0)&(Ap05==1)]<-1
#conditional: Ap is not sig, but RMSp is
CRMSp05<-matrix(0,nrow=loops,ncol=(nincreases+1)); CRMSp05[(RMSboth05==1)&(Ap05==0)]<-1
#collects count of significance as a % of significant tests for RM interaction, T2 simple
effect, and ANCOVA group significance
```

```
p05<-matrix((colMeans(RMp05)),nrow=1, ncol=(nincreases+1), byrow=TRUE);
p05<-rbind(p05,(colMeans(RMSboth05)),(colMeans(Ap05)),colMeans(CAp05),colMeans(CRMSp05))
#table with difference
pdifwin<-matrix((p05[3,]-p05[2,]),nrow=1,ncol=(nincreases+1));
p05<-rbind(p05,pdifwin)


#name rows and columns
rownames(p05)<-
c("RMANOVA_0.05","RM_any_simple_effect","ANCOVA_0.05","Ap_sig_but_not_RMSp","RMSp_sig_but
_not_Ap","ANCOVA_minus_RM");
colnames(p05)<-seq(from=10, to=(nincreases*10+10), by = 10)


finish<-Sys.time() #timestamp to know the time needed
#the simulation required a
(finish-start)
#for a total of
(loops*(nincreases+1))
#loops#p1;#;p01;p001;p0001;
p05;#pdif;pdifwin
```

Table 1: Summary of simulations in the litterature

article	allocation				parameters								outcome***	
	random	pre-test assigned	<- mix ->	pre-existing	effect size (pre-post) r	effect size (pre-post) β (G with outcome)	effect size (pre-post) d	n (total)	repetitions	baseline imbalance	reliability	normality	ρ(X1ΔX)	Correct rejection of H0	Type I error
Estimating Sample Sizes for Repeated Measurement Designs (1994), John E. Overall and Suzanne R. Doyle	x				0.3		0.5		10 000					125/63*	
	x				0.5		0.5		10 000					53/69*	
	x				0.7		0.5		10 000					40/27*	
	x				0.3		0.2		formula					587/398*	
	x				0.5		0.2		formula					418/321*	
	x				0.7		0.2		formula					250/214*	
	x				0.3		0.5		formula					94/63*	
	x				0.5		0.5		formula					64/51*	
	x				0.7		0.5		formula					40/34*	
	x				0.3		0.8		formula					37/25*	
	x				0.5		0.8		formula					26/20*	
	x				0.7		0.8		formula					16/14*	
ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies (2006), Gerard J.P. Van Breukelen **		x			0.52			20	1						Δ falsey rejects H0, ANCOVA does not
				x			0	20	1	exp. group +10					ANCOVA falsely rejects H0
Comparing groups in a before-after design (2006), Daniel B Wright		x				0, .5, 1, 2		100	1000					t-test has more power	t-test biased strongly as error increased
			x	.		0, .5, 1, 2		100	1000					t-test is underpowered	ANCOVA biased, but error not sig. Increased
			x	.		fixed between groups		100	1000					same results as above	same results as above
	x					0, .5, 1, 2		100	1000					ANCOVA is more powerful	no bias for t-test or ANCOVA
		x	x			0, .5, 1, 2		100	10 000					neither are appropraite, high Type I error rate	
		x				0, .5, 1, 2		100	1000					t-test biased, ANCOVA unbiased with and without interaction	
	x					0, .5, 1, 2		100	1000					both unbiased, always	
			x	.		0, .5, 1, 2		100	1000					with slopes fixed, t-test unbiased, otherwise, both biased	
A Simulation Study on the Performance of the Simple Difference and Covariance-Adjusted Scores in Randomized Experimental Designs (2011), Yaacov Petscher & Christopher Schatschneider	x				.2, .6, .8			40, 60, 100, 400, 1000	1000			skew = ±.5; 1	0.3	equal at .48	state that .05 is respected overall
	x				.4, .6, .8			40, 60, 100, 400, 1000	1000			skew = ±.5; 1	0	RM .59 ANCOVA .61	
	x				.6, .8			40, 60, 100, 400, 1000	1000			skew = ±.5; 1	-0.3	RM .60 ANCOVA .69	
	x								1000			skew = ±.5; 1	0.3	skew had very little effect	
	x								1000			skew = ±.5; 1	0	skew had very little effect	
	x								1000			skew = ±.5; 1	-0.3	skew had very little effect	
	x							40, 60, 100, 400, 1000	1000				0.3	no difference	
	x							40, 60, 100, 400, 1000	1000				0	as N grows, ANCOVA's advantage shrinks	
	x							40, 60, 100, 400, 1000	1000				-0.3	as N grows, ANCOVA's advantage shrinks	
	x				.2, .6, .8				1000				0.3	no difference	
	x				.4, .6, .8				1000				0	gap shrinks with r, ANCOVA wins	
	x				.6, .8				1000				-0.3	gap srhinks but remains important, ANCOVA wins	
A Monte Carlo Comparison Study of the Power of the Analysis of Covariance, Simple Difference, and Residual Change Scores in Testing Two-Wave Data (2012), Yasemin Kisbu-Sakarya, David P. MacKinnon, Leona S. Aiken	x				.3, .5, .7, 1	±0, .14, .39, .59		100	1000	none	.5, .8, 1			never reached p=.8	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		200	1000	none	.5, .8			never reached p=.8	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		200	1000	none	1			p-> .8 for large ES	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		400	1000	none	.5, .8			p-> .8 for large ES	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		400	1000	none	1			p-> .8 for medium ES	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		1000	1000	none	.5, .8, 1			p-> .8 for medium ES	~.04-.07
	x				.3, .5, .7, 1	±0, .14, .39, .59		100, 200, 400, 1000	1000	sm & lrg	.5, .8			ANCOVA wins when $r(\beta_G \Delta x)$ is positive	as reliability and imbalance worse, Type I error up for ANCOVA
	x				.3, .5, .7, 1	±0, .14, .39, .59		100, 200, 400, 1000	1000	sm & lrg	1				~0.05

*n to reach p=.8 for RM/ANCOVA

** Original sample has a fixed 180 total, 148 complete (pre: 88/92; post: 68/80); all "simulations" sampled from this

***results did not necessarily talk of correct rejection of H0 or Type I error; they were interpreted as best as possible in these terms to allow some semblece of comparison of the results across two decades of research