Fall 11-30-2018

Using Bayesian Multilevel Models to Control for Multiplicity among Means

Michael J. Zweifel
University of Nebraska-Lincoln, zweifemj@huskers.unl.edu

USING BAYESIAN MULTILEVEL MODELS TO CONTROL FOR MULTIPLICITY AMONG MEANS

by

Michael Zweifel

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

(Quantitative, Qualitative, and Mixed Methods)

Under the Supervision of Professor Rafael J. De Ayala

Lincoln, Nebraska

November, 2018

USING BAYESIAN MULTILEVEL MODELS TO CONTROL FOR MULTIPLICITY AMONG MEANS

Michael Zweifel, Ph.D.

University of Nebraska, 2018

Adviser: Rafael J. De Ayala

It is well known that the Type I error rate will exceed α when multiple hypothesis tests are conducted simultaneously. This is known as Type I error inflation. The probability of committing a Type I error grows monotonically as the number as the number of hypothesis being tested increases. A class of methods, known as multiple comparison procedures, has been developed to combat this issue. However, in turn for maintaining the Type I error rate below α, multiple comparison procedures sacrifice power to correctly reject false hypotheses. The loss of power is exacerbated when variance heterogeneity is present.

In the case of making multiple comparisons among means, a possible alternative to multiple comparison procedures is to use Bayesian multilevel models to control for Type I error inflation. Bayesian multilevel models reduce the risk of committing a Type I error by shrinking all means towards the grand mean, in turn, making it more difficult to declare any mean significantly different from one another.

To compare the performance of multiple comparison procedures and Bayesian multilevel models, a Monte Carlo simulation study, in which the number of hypotheses and variance heterogeneity was manipulated, was conducted. The results indicated that the Bayesian multilevel models maintain the Type I error rate at α and display greater power than the traditional methods when a large number of hypotheses are tested. When

the number of hypotheses tested were small, the Bayesian models were not able to maintain strong control of the Type I error rate.

Table of Contents

List of Tables

List of Figures

CHAPTER I. INTRODUCTION

Background

In educational policymaking, it is often necessary to make decisions about the performance of schools, teachers and programs. Among other possible outcomes, these decisions may influence whether a school is administered sanctions, a teacher is recommended for promotion, or the continuation of resources towards an educational policy. A variety of standards are used to make such high stakes decisions; however, one measure is to evaluate performance relative to a criterion via hypothesis testing. For instance, the performance of a school might be evaluated by testing whether that school's average score on a standardized instrument is statistically different from some, pre-determined benchmark. As another example, a hospital's success at performing ome operation might be judged against the average success rate for all hospitals in a state. In such a scenario, a hypothesis test would be conducted for each hospital under evaluation. Under the frequentist paradigm to statistics, the probability of committing a Type I error is inflated when multiple hypotheses are tested simultaneously. This inflation issue is called the *multiplicity* or *multiple comparisons problem*. In the context of educational policy decisions, an increased probability of arriving at an incorrect conclusion could have undesirable consequences.

A set of methods, known as multiple comparison procedures (MCPs), have been developed to control the Type I error rate at an upper limit of α when multiple hypotheses are tested simultaneously. This property is known as strong

control of the Type I error rate. Traditional MCPs typically treat the sample means under consideration as estimates of a *fixed* population mean. The performance of MCPs are primarily judged on two qualities: The ability to maintain strong control of the Type I error rate and the power to correctly detect false null hypotheses (Ramsey, 1981). Two popular classes of MCPs are methods that adjust the *p*-values generated from statistical tests and methods that control for multiplicity by utilizing the studentized range distribution. (Shaffer, 1995). Within these two classes, MCPs have been developed that guarantee strong control of the Type I error rate. Both classes of MCPs make distributional assumptions about the data, such as normality and variance homogeneity.

Unfortunately, in exchange for maintaining strong control of the Type I error rate at α, MCPs sacrifice significant power to detect true significant hypotheses. Further, as the number of hypotheses being tested increases, the probability of committing a Type I error increases as well. To combat this, MCPs become increasingly conservative as the number of hypotheses rise making it more difficult to correctly detect false hypotheses. This loss of power is further exacerbated when the distributional assumptions of the MCPS are violated (Shaffer, 1995). In particular, the presence of variance heterogeneity has been shown to significantly reduce the power of MCPs (Games & Howell, 1976; Kromrey & La Rocca, 1995; Shaffer, 1995). It would be desirable if an alternative method existed which exhibited strong control of the Type I error rate while maintaining the power to correctly reject false null hypotheses in the face of such circumstances. Multilevel models (MLMs) may provide such an alternative.

MLMs specify two or more, hierarchical levels of relationships among parameters (Greenland, 2000). For example, student performance, termed the level-one unit, is nested within the school the student attends, the level-two unit. Unlike fixed effects estimation, MLM is a *random effects* model, which assumes that the higher-level units are drawn at random from some larger population. Inclusions of random effects are typically treated as nuisance parameters intended to account for unexplained variance in lower level units (Raudenbush & Bryk, 2002). However, educational researchers have used MLMs to test hypotheses about level-two means, such as using student level data to make inferences about schools (Ohlessen, Sharples, & Spiegelhalter, 2007).

MLMs provide some control for multiplicity by shrinking the level-two group mean estimates towards the aggregate mean (Gelman, Hill, & Yajima, 2012; Normand, Glickman, & Gastonis, 1997). This results in level-two mean estimates that are closer to one another as compared to their fixed effects counterparts. Moreover, the degree of reduction to the aggregate mean is influenced by the within and between level variances, not the level-two sample size. This suggests that when using MLMs to conduct hypothesis tests about the level-two means, increasing the number of tested hypotheses will not affect the power to detect true statistically significant results. However, frequentist MLMs are not used to control for multiplicity because it is not possible to evaluate the Type I error rate of these models via simulation study. The reason for this is when data are simulated so that no variance exists between the level-two units, the

frequentist multilvel models reduces to an ordinary least squares regression model in which the random effects are identical for all level-two units.

Adopting a Bayesian approach to multilevel models could avoid this problem and is discussed in more detail below. The Bayesian paradigm may offer additional adjustments that allow researchers to ensure strong control of the Type I error rate and to directly incorporate variance heterogeneity into their model. The Bayesian approach to statistics differs from the frequentist approach in that all model parameters are assigned a probability distribution rather than being assumed to have some true, fixed value (Lynch, 2007). Parameter estimates are typically summarized by the descriptive statistics of a distribution (called the posterior distribution) which is formed as the product of the likelihood estimate of the parameter of interest and a hypothesized sample space in which the parameter may lie (called the prior distribution). The prior distribution may be chosen based upon previous knowledge or theory such that the prior distribution encapsulates all possible values of the true parameter (Kaplan, 2014). The first case is referred to as an informative prior, whereas the second case is referred to as an uninformative prior. This study will use uninformative priors for all parameters, except when noted otherwise. Bayesian analysis has several advantages over frequentist methods, among which are the ability to incorporate the results of previous research into the prior distributions and increased flexibility in modeling MLMs.

The concept of Type I error inflation does not exist in the Bayesian framework because there is no assumption that the true population parameter is

fixed (Freedman, 1996). However, it is still possible to reach incorrect conclusions. Further, when uninformative prior distributions are used, Bayesian inferences approximate frequentist estimates. As a result, Bayesian models are not guaranteed to maintain the Type I error rate at a value pre-determined by the researcher.

The Bayesian MLM (Gelman et al., 2012) with uninformative priors is very similar to the frequentist MLM; however, there is one crucial difference. This difference is that the Bayesian MLM assumes that the variance components in the MLM are drawn from a random distribution and, consequently, are assigned their own prior distribution. By placing a prior distribution on the between groups variance components, it is possible to evaluate the empirical Type I error rate of MLMs. Even if the between group variance is specified to arise from a distribution with a mean of zero, there is still some unique error variance incorporated into the estimated random effect of a level-two unit. Beyond that, the frequentist MLM and unadjusted Bayesian MLM should provide similar benefits for controlling multiplicity.

An alternative method of specifying the Bayesian MLM is to assign a parameter that denotes the difference between any mean and a criterion, this difference is represented as δ_q. This difference parameter is assigned a mixture prior distribution with a hyper-parameter that signifies the probability that the difference between a mean and the criterion is 0 (Li & Shang, 2015; Nashimoto & Wright, 2008; Shang, Cavanaugh, & Wright, 2008; Shang, 2011). If this parameter indicates that the difference between the mean and a criterion is zero,

then the difference prior distribution is assigned to be a point mass prior with its entire mass at zero. Otherwise, δ_q is assigned a continuous distribution (Li & Shang, 2015; Nashimoto & Wright, 2008).

The Bayesian MLMs described so far do not account for variance heterogeneity among the level-two units beyond the adjustment all MLMs make by shifting means with large variances closer to the grand mean. Recalling that Bayesian analysis assigns all model parameters a random distribution, the Bayesian MLMs described thus far assign a common prior distribution to the variance of each level-two mean. In order to account for variance heterogeneity, these models may be expanded so that the variance of each level-two mean is assigned its own unique prior distribution. The result of this is that variance of each level-two mean is allowed to vary depending on the sample data (Nashimoto & Wright, 2008). The variance for each level-two unit, alternative referred to as the within groups variance, is denoted by σ_{ij}^2. These unique prior distributions are chosen by using the estimate of each level-two sample variance to inform the variance of the prior distribution. In this way, differences in variability between level-two units are directly modeled. These semi-informative prior distributions for the within group variance components may be applied to both methods of Bayesian MLMs. Like the frequentist MLMs, there is a lack of empirical research on the operating characteristics of these Bayesian methods when used as corrections for multiplicity.

Purpose

The purpose of this study was to evaluate the Type I error control of four MLM approaches (Bayesian MLM, Bayesian MLM with a difference parameter, both Bayesian approaches to MLMs with unique prior distributions for σ_j^2) when testing whether several level-two means differ significantly from a criterion. The Type I error rate of these procedures was also compared to the Type I error rate of two traditional MCPs (Hochberg and Tukey's HSD procedures). Additionally, this study evaluated the extent to which these six methods correctly detect when the level-two means differed significantly from a criterion, especially when a large number of comparisons were made and variance heterogeneity was present among the level-two means.

A Monte Carlo simulation study was conducted to evaluate the Type I error rate and power of these procedures. Data were simulated from a three-level multilevel model. While the primary focus of this study is to examine level-two means, a third hierarchical level was included in the data generation process to simulate the scenario in which unexplained covariance between level-two units was present. The factors that were manipulated were the mean difference of the level-two units from the criterion, the presence of level-one variance heterogeneity, the number of level-two units, the amount of variance between level-two units due to level-three variability, and the amount of variance in the level-one units that was due to variance in the level-two units. Table 1 summarizes the levels of the independent variables.

Table 1. *Independent Variables*

Factor	Levels
Within Group Variances	All σ^2_{ijk} =1 50% of σ^2_{ijk} =.5 & 50% of σ^2_{ijk} =1.5
Level-two Sample Size	N_j = 8 N_j = 20 N_j = 40
Effect Size	β_{ijk} = 0 β_{ijk} = .2 β_{ijk} = .5
Level-two *ICC*	$ICC2$ = 0 $ICC2$ = 0.15 $ICC2$ = 0.25
Level-three *ICC*	$ICC3$ = 0 $ICC3$ = 0.1
Procedures	Hochberg's MCP Tukey's MCP Bayesian MLM Bayesian δ MLM Variance Informed Bayesian MLM Variance Informed Bayesian δ MLM

The levels of σ^2_{ijk} were chosen to ensure that a sufficient number of level-two units exhibited variance heterogeneity. Because previous research (e.g., Games & Howell, 1976; Kromrey & La Rocca, 1995) has demonstrated that even small amounts of variance heterogeneity results in noticeable power loss, a moderate degree of variance heterogeneity was selected. The level-two sample sizes were chosen to facilitate the variance heterogeneity conditions and to allow for an increasing number of comparisons among the level-two units. Additionally, research has demonstrated that MLMs may not produce reliable estimates when the number of level-two units are small (Raudenbush & Bryk,

2002). The effect sizes were chosen in accordance with Cohen's suggestions for small and medium effects (1988). The levels of the intraclass correlation coefficient, the *ICC*, were chosen to align with values of the *ICC* that were commonly reported in educational research (Hedges & Hedberg, 2007a; Hedges & Hedberg, 2007b).

The dependent variables in this study were the Type I error rate and power of each procedure. The Type I error rate for each procedure was evaluated under the condition in which the effect size equal to zero, resulting in 54 experimental cells. The power of each procedure was evaluated under the condition in which the effect size was not equal to zero, resulting in 60 experimental cells.

Organization

The organization of the remainder of the text is as follows. Chapter Two provides an overview of multiple testing, traditional approaches to controlling for multiplicity, and limitations to traditional MCPs. Additionally, frequentist and Bayesian MLMs are introduced and the use of Bayesian models as MCPs is explored. A literature review of research in this field is provided. Chapter Three describes the methods used to conduct this study. Chapter Four presents the results of the study. Chapter Five discusses the implications of the results in addition to describing limitations and future directions.

Significance

The present study is novel in that it provides an empirical examination of the operating characteristics of Bayesian MLMs when used as a correction for Type I error inflation and compares their performance to traditional MCPs. This

study provides insight as to whether MLMs may be a viable alternative to traditional MCPs, particularly when a large number of hypotheses are being tested and variance heterogeneity is present.

CHAPTER II. LITERATURE REVIEW

When statistical tests are used as a part of the educational policy decision-making process, it is common for investigators to test several hypotheses simultaneously. In the frequentist approach to statistics, doing so results in an inflation of the Type I error rate, which corresponds to a higher probability of making an incorrect decision. Several procedures have been developed to control the Type I error rate at or below α. However, these procedures are known to suffer a severe loss of power when variance heterogeneity is present among the level-two units and a large number of hypotheses are being tested. MLMs may provide an alternative for controlling multiplicity. MLMs provide some inherent control for Type I error rate inflation by shifting level-two means closer to the aggregate mean. In the case in which no level-two mean differs significantly from the criterion, MLMs would estimate the grand mean as the criterion and the level-two mean estimates would be drawn towards the criterion. This, consequently, would make it more difficult to commit a Type I error. Further, estimates of level-two means become more accurate as the number of level-two units increase. Adopting a Bayesian approach to MLMs allows the researcher to incorporate the presence of variance heterogeneity directly into the model.

The following chapter discusses several topics related to multiplicity. The chapter begins by discussing Type I error rate inflation and traditional procedures used to control for Type I error rate inflation. Next, the use of MLMs as a control for multiplicity is introduced, along with a method for testing hypotheses about level-two means. The Bayesian approach to statistics is then introduced, along

with how the Bayesian paradigm may be applied to MLMs to better control for Type I error rate inflation, particularly when variance heterogeneity is present.

Hypothesis Testing

The field of statistics involves describing and making inferences about a population parameter. The goal of statistical inference is to describe unknown parameters using observed data (Kaplan, 2014). In an ideal world, information would be gathered from all members of the population and the characteristics of the parameters of interest would be known. However, this is often not feasible for many reasons, such as resource limitations or lack of access to all members of the population. Instead, researchers often select a sample from the population and use sample statistics to estimate the relevant population parameters.

As a motivating example for the remainder of the paper, consider the scenario in which n_i students exclusively attend one of J high schools. A researcher wishes to compare these schools' average academic achievement to the average academic achievement of the population of schools in the nation. The ostensible purpose of such an endeavor may be to identify those schools that are under or over performing relative to their peer schools. Note that this scenario may easily be extended to the situations in which a researcher is comparing the academic achievement of several schools against one another, known as the pairwise comparisons case, or evaluating several schools' academic achievement against a control school, known as the multiple one case.

The motivation for identifying these schools may be to reward those schools designated as high achieving and to provide additional aid or

improvement to those schools designated as low achieving. The national average academic achievement will be referred to as the criterion. Assume that academic achievement is operationalized as a normal, continuous outcome and has been grand mean centered such that the average national academic achievement is zero. In this paper, schools will be referred to alternatively as the independent variable or the level-two unit and students will be referred to as the level-one unit. This may be represented in the one-way ANOVA paradigm as:

$$y_{ij} = \mu_j + e_{ij}, \quad (1)$$

where y_{ij} is academic achievement score for student i in school j ($j = 1 \ldots J$), μ_j is the mean academic performance in school j, and e_{ij} is an error term distributed as $N(0,\sigma^2)$. This scenario may also be expressed identically as an ordinary least squares linear regression model with a categorical predictor:

$$y_{ij} = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + e_{ij}, \quad (2)$$

Where β_0 is the mean of the school assigned as the reference school (also referred to as the intercept), β_j corresponds to the mean difference in academic achievement between the remaining *J-1* schools and the reference school and each x_j is a binary variable indicating which school is being examined. Further, Equation 1 is analogous to an ordinary linear least squares model with school as a categorical predictor and with the intercept removed. Because these three models produce identical results, the ANOVA and regression models will be used interchangeably in this paper.

Let us assume that an individual school's influence on a student's achievement score is a *fixed effect*. The effect of schools is said to be fixed when

the schools are specifically selected by the experimenter. As a result, generalizations are restricted to only those schools included in the model. Additionally, the fixed effect estimate of an individual school's performance is assumed to be completely independent from the effect of any other school in the sample.

Making multiple comparisons among level-two means by using level-one data has real world parallels. For instance, the National Center for Education Statistics (1997) produced a report which ranked each US state by mean mathematics performance in mathematics on the National Assessment of Educational Progress test. Federal educational initiatives such as the *No Child Left behind Act of 2001* and the *Elementary and Secondary Education Act* have inspired researchers to investigate how student level data, such as standardized test scores, may be used to infer knowledge on the performance of schools and teachers (Shaw, 2012). The United Kingdom's Parents' charter requires the department of education to publish rankings of secondary schools using the results of the General Certificate of Secondary Education exam (Goldstein & Thomas, 1996; Leckie & Goldstein, 2009). The purpose of publishing these results is to provide parents with a metric to use when choosing which school for their children to attend. When considering teacher evaluation, value added models provide scores on teacher effectiveness based upon the achievement of their students (Schochet & Chiang, 2013).

Inferences drawn from comparisons between level-two units is not limited to the field of education. In medical research, biostatisticians may perform

subgroup analysis to evaluate differences in treatment effects for groups of patients who differ on some baseline characteristic (Wang, Lagakos, Ware, Hunter, & Drazen, 2007). Similar to comparing schools using student level data, it is often desirable to evaluate hospital performance using patient level data, such as the mortality risk among myocardial infarction patients (Austin, Naylor, & Tu, 2001). As another example, genetic association studies are often concerned with using data from multiple genetic variations nested within candidate genes in order to detect genes that are more likely to be associated with a disease (Yi, Xu, Lou, & Mallick, 2014).

To evaluate the question of whether a school's academic achievement differs significantly from the criterion, again defined as the grand mean centered national achievement average, the researcher specifies two hypotheses for each of the *J* schools included in the study. The null hypotheses (H_0) for any school states that the mean academic performance for that school does not differ significantly from the criterion. This may be written as $H_0: \mu_j = 0$. The alternative hypothesis (H_1) states that the null hypothesis is not true, in this case meaning that the school under consideration does differ in academic achievement from the criterion. This may be written as $H_1: \mu_j \neq 0$. Null and alternative hypotheses are stated for every school under consideration. The researcher then collects statistical evidence in order to make an inference about whether a given school's true population academic achievement is different from the criterion.

Frequentist Paradigm

What, then, constitutes enough statistical evidence to reject the null hypothesis and declare that a school's academic performance differs statistically from the national average? The frequentist paradigm in statistics provides one philosophy for answering the above question. This is done by constructing the distribution of a sample statistic. In our example, we are concerned with the distribution of sample means. Given that level-one units are sampled at random from a population, the distribution of sample means is composed of the means taken from every possible combination of level-one units of sample size *n*. Constructing the distribution of sample means assumes that it is possible to sample all possible permutations of level-one units, which is not realistic if the population size is even moderately large. However, the central limit theorem states that for samples of sufficient size (30 is a generally accepted as a rule of thumb; Gravetter & Wallnau, 2017) the distribution of sample means will be normally distributed; the distribution of sample means will have a mean equal to the mean of the total population; the standard deviation of the distribution of sample means will be equal to the standard deviation of the population divided by the square root of the sample size (Gravetter & Wallnau, 2017).

The central limit theorem allows the researcher to calculate the probability that the absolute value of a randomly selected sample mean is greater than the absolute value of the population mean due to chance alone; this probability is called the *p*-value. Stated more formally, the *p*-value is the probability of observing a sample statistic that is equal to or more extreme than the parameter's

value given the null hypothesis (Wasserstein & Lazar, 2016). If this probability is sufficiently small, then there is evidence that the population parameter is most likely something other than the value specified in the null hypothesis. Conversely, a large *p*-value conveys that there is a lack of evidence that the population parameter is different from the value specified in the null hypothesis (Wasserstein & Lazar, 2016). Researchers set the criterion for what defines a small probability for an observed sample statistic. This probability value, α, is chosen *a priori*. Researchers reject the null hypothesis when the obtained *p*-value is less than or equal to α and retain the null hypothesis otherwise. Most commonly α is set to .05; this will be the value used in this paper. Equivalently, researchers may construct a confidence interval around the sample estimate. A 95% confidence interval around the sample statistic signifies that if an infinite number of independent samples of a given size were drawn from the population of interest, then the probability is .95 that a randomly sampled confidence interval will contain the population parameter (Gravetter & Wallnau, 2017). With respect to testing the null hypothesis, confidence intervals that do not contain the hypothesized parameter value provide evidence that the null hypothesis is not true.

There are several implications of the frequentist approach to statistics. First, the frequentist paradigm makes the assumption that the parameter of interest has a single, true value that is unknown to the researcher. Second, frequentist hypothesis testing only allows inferences about the null hypothesis. Perhaps most

importantly, the frequentist approach to statistics is based upon the premise of a population being sampled an infinite number of times.

Type I/Type II error.

When testing hypotheses, it is possible to make incorrect inferences. Specifically, a researcher may make a Type I error by rejecting a null hypothesis that is true or a Type II error by failing to reject a null hypothesis is false. When testing a single hypothesis, the Type I error rate will be equal to α. In this way, the researcher is able to select the tolerable Type I error risk. It is generally not desirable to specify the tolerable Type II error risk *a priori*. Table 2 provides a summary of the two errors.

Table 2. *Type I and II Errors*

		Decision	
		Retain H_0	Reject H_0
Reality	H_0 True	Correct Decision	Type I Error
	H_0 False	Type II Error	Correct Decision

Type I errors occur because the distribution of sample means under the null hypothesis contains legitimate values that have a probability of less than α of being sampled (Gravetter & Wallnau, 2017). The probability of randomly selecting a given sample mean is determined by its distance from the mean specified under the null hypothesis divided by the variability of the distribution of sample means. When the probability of randomly sampling these values is small, there appears to be evidence that the mean under the null hypothesis is incorrect and the researcher wrongly concludes the null hypothesis is false.

Type II errors occur because a portion of distribution of sample means around the true population mean overlaps with the critical region of the distribution of sample means as specified by the null hypothesis and α. This leads researchers to incorrectly conclude there is no effect when in fact the null hypothesis is false. Type II errors often occur because the statistical test does not contain enough power to detect a true, significant effect (Gravetter & Wallnau, 2017).

Researchers are generally more concerned with the risk of committing Type I errors than the risk of committing Type II errors (Ludbrook, 1998). Committing a Type I error may lead to the spread of misinformation, incorrect policy implementations, or the adoption of a potentially harmful treatment (Ludbrook, 1991). Type II errors, on the other hand, may lead a researcher to fail to implement a useful treatment. The severity of committing a Type I error as compared to committing a Type II error can be seen in the motivating example. Committing a Type I error would result in a school being incorrectly identified either over or under performing. Hypothetically, suppose schools identified as underperforming are subject to a loss of federal funding. A Type I error then would result in the incorrect decision to reduce funding in the school that the Type I error was committed against. A Type II error would result in the declaration that a school's academic achievement did not differ significantly from the criterion when, in fact, it did. Committing a Type II error, on the other hand, would result in no appreciable consequence because the school's performance would not be distinguishable from the criterion.

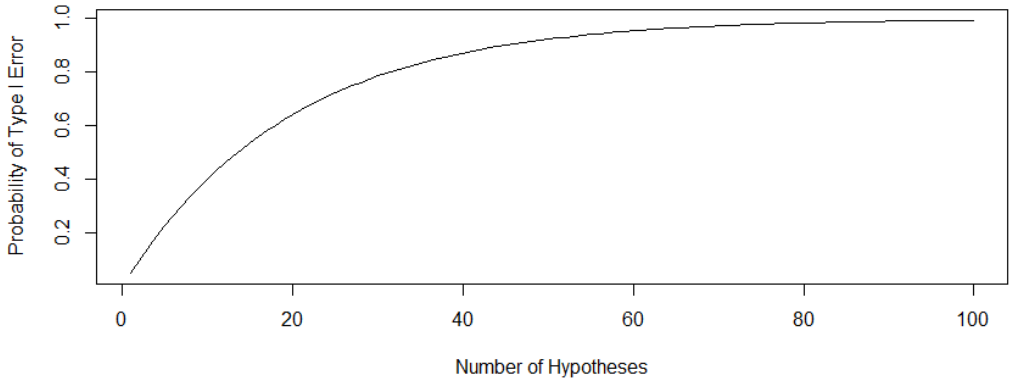
Multiplicity.

In the frequentist framework of statistical analysis, when multiple hypotheses are tested on a single data set the Type I error rate will not be maintained at α. Instead, the probability of committing a Type I error is inflated. Recall that for a single hypothesis test, the researcher has a probability of α for committing a Type I error. If another hypothesis is tested, the probability of committing at least one Type I error between the two tests is compounded. Assuming that the hypotheses being tested are independent of one another, the probability of observing one or more Type I errors may be calculated as:

$$1-(1-\alpha)^C, \tag{3}$$

where C is the number of hypotheses to be tested (Abdi, 2010) and α is the (common) significance level used for the C tests. As an illustration using the ongoing example, suppose separate hypothesis tests are conducted to identify whether the academic achievement of four schools differs significantly from the criterion. If α is set to .05, then Equation 3 yields a probability value of .186 for committing a Type I error when, in reality, none of the four schools differ from the criterion. This phenomenon is known to as the *multiplicity* or multiple comparisons problem. Figure 1 illustrates how the probability of committing at least one Type I error increases with the number of tested hypotheses.

Figure 1. *Relationship between Number of Hypothesis & the Type I Error Rate*



When hypotheses are not independent of one another, as in the case of making all pairwise comparisons, it is not possible to calculate the extent of the Type I error rate inflation *a priori* due to the fact that the degree of dependency is not known beforehand. However, research has provided evidence that dependent hypotheses exhibit a slightly lower degree of Type I error rate inflation (Fink, McConnell, & Vollmer, 2014).

Implications.

In educational research, it is common for several hypotheses to be tested simultaneously. Gelman et al. (2012) list examples such as examining differences in demographic information across multiple schools, counties, and states, searching for differences among subgroups of a population, or evaluating the effect of an intervention on several outcomes. As a result, multiplicity presents a serious concern to applied researchers.

There is some debate as to when one should be concerned about Type I error rate inflation. Ryan (1959) listed five situations in which it may be appropriate to control for multiplicity. The first situation is when more than two

groups are being compared to one another. The second situation is when determining whether significant correlations between three or more variables exist. The third situation is when the researcher is attempting to determine which main effects and interactions are significant in a factorial analysis of variance (*ANOVA*) design. The fourth situation is when a researcher tests the significance of the same experiment over several different independent samples. The final situation that Ryan describes is when several different measures for evaluating a variable are compared. Additionally, several authors (Bender & Lange, 2001; Dunnett & Tamhane, 1992) have argued that it is necessary to control for inflated Type I error rate in confirmatory studies and whenever multiple hypothesis have to be synthesized to a single conclusion.

In contrast, some researchers have argued against controlling for multiplicity (deCani, 1984; Gelman et al., 2012; O'Keefe, 2003). Several authors (Ludbrook, 1991; Seaman, Levin, & Serlin, 1991) have posited that Type I error rate inflation is not a concern in exploratory settings because the researcher is approaching the data from a uninformed perspective and wishes to uncover interesting relationships within the data. O'Keefe (2003) noted that even if one justifies controlling for multiplicity in an experimental setting, the application of Type I error rate adjustment is inconsistent in practice. O'Keefe (2003) further states that while it is common for researchers to control for multiplicity in the *ANOVA* setting, it is less common for researchers to correct for Type I error inflation when analyzing regression models with multiple predictors or when examining several bivariate correlation coefficients.

Others have argued that it is not necessary to use multiple comparisons when independent, planned, or *a priori*, comparisons are used (Hays, 1994; Keppel & Wickens, 2004; Ludbrook, 1991). In order to test these planned contrasts, a set of weights are applied to each mean. The values of the weights are chosen such that hypothesized comparisons of interest can be made among the means. The sum of the weights multiplied by the group means is the estimated comparison value. An *F* statistic may be constructed from the estimated comparison value and used to provide evidence that there is a significant difference between the contrasted groups. Assuming orthogonal contrasts, the Type I error rate will be maintained at α for all contrasting hypotheses. Comparisons are said to be orthogonal of one another if, assuming equal sample sizes, the sum of the products of weights assigned to each comparison is equal to zero (Hays, 1994). A drawback to planned comparisons is that the number of orthogonal comparisons that can be made is limited to the number of level-two units minus one (Toothaker, 1991). Non-orthogonal *a priori* contrasts, such as in the case of all pairwise comparisons, do not maintain the Type I error rate at α, because the comparison being made is based on redundant information already gathered from a previous comparison (Hays, 1994).

By far the most common argument against controlling for multiplicity is that doing so sacrifices too much power to correctly reject the null hypothesis (Benjamini & Hochberg, 1995; deCani, 1984; Dunnett & Tamhane, 1992; O'Keefe, 2003). While the loss in power from using MCPs can be substantial, researchers have generally agreed that maintaining the Type I error rate at α is of

greater concern than maintaining power (Tollenaar & Mooijaart, 2003). As a result, it is typically considered necessary to account for multiplicity when several hypotheses are tested simultaneously.

Multiple Comparison Procedures

Accepting that it is generally necessary to control for multiplicity when several hypotheses are tested on the same data set some action must be taken. MCPs have been specifically developed to address multiplicity. Most MCPs increase the threshold(s) necessary to declare that a test's result is significant while attempting to retain the greatest amount of power to detect true differences. The effectiveness of MCPs is judged primarily on two qualities: The ability to maintain the Type I error rate at or below α, and the ability to correctly detect false null hypotheses, referred to as the power of the MCP (Ramsey, 1981). Before describing the properties of effective MCPs a more thorough discussion of the Type I error rate is needed.

Type I error definitions.

There are a number of possible definitions of the Type I error rate when testing multiple hypotheses. As a result, there is some debate in the literature regarding which definition of the Type I error rate is most appropriate (Ramsey, 1981). To define the Type I error rate, the researcher must first determine what constitutes the family of hypotheses to be tested. The family of hypotheses informs the extent of the Type I error rate inflation. The most general definition of a family of hypotheses is the set of hypotheses that the researcher evaluates during an experiment or study (Games & Howell, 1976; Ludbrook, 1998; Shaffer,

1995). Ultimately, the make-up of the corresponding family of tests depends on the purposes of the study and the research questions being asked (Ludbrook, 1998; Shaffer, 1995).

Let C index the number of null hypotheses in a family of tests and m_0 denote the number of retained null hypotheses. Table 2 may be redrawn for multiple hypothesis tests as:

Table 3. *Type I and II Errors Notation*

H_0	Retained	Rejected	Total
True	U	V	C_0
False	T	S	$C - C_0$
Total	W	R	C

Several definitions of the Type I error rate may be constructed using Table 3 (Bretz, Hothorn, & Westfall, 2011). Three of the most popular definitions are the *per-comparison* (PCE) error rate, the *familywise* error rate (FWE), and the *false discovery rate* (FDR).

The per-comparison error rate is the simplest definition of the Type I error rate. The per-comparison error rate is also known as the *comparisonwise* error rate, the *individual level*, or the *individual* error rate (Bender & Lange, 2001). The per-comparison error rate is the expected proportion of Type I errors among all hypotheses being independently and separately tested, and may be written as:

$$\alpha_{PCE} = \frac{E(V)}{C}, \tag{4}$$

where E(V) is the expected number of Type I errors in a family of tests. The per-comparison error rate is equal to α for each test and procedures that control for the per-comparison error rate essentially ignore the effects of multiplicity. As a

result, this definition of the Type I error rate is not useful when testing multiple hypotheses.

The familywise error rate is the probability of committing at least one Type I error among a family of tests (Games & Howell, 1976; Ryan, 1959) and is written as:

$$\alpha_{FWE} = P(V > 0), \quad (5)$$

where $P(V>0)$ is the probability that at least one Type I error occurs in a family of tests. This is the most commonly used definition of the Type I error rate and a number of MCPs have been developed to control the familywise error rate at or below the value of α_{FWE} selected *a priori* by the researcher (Bretz et al., 2011). Given independent tests, the familywise Type I error rate may be determined by Equation 3.

Alternatively, the false discovery rate is the expected proportion of falsely rejected hypotheses among the total number of rejected hypotheses:

$$FDR = E(\frac{V}{R}). \quad (6)$$

If no hypotheses are rejected, then R is equal to zero and the FDR is set equal to zero (Benjamini & Hochberg, 1995; Bretz et al., 2011). The false discovery rate is equal to the familywise error rate when all hypotheses are true but smaller than it when at least one hypothesis is false (Benjamini & Hochberg, 1995). Any MCP that controls for the familywise error rate will also control for the false discovery rate, but the converse is not necessarily true (Bretz et al., 2011)

It appears as if the familywise error rate has become the most popular definition of the Type I error rate when making multiple comparisons (Brown &

Russell, 1997; Einot & Gabriel, 1975; Lehmann & Romano, 2005; Ludbrook, 1998). The familywise error rate definition has often been used when the size of the family of hypotheses is small to moderate or when strong support for rejecting the null hypothesis is required (Bretz et al., 2011). As a result, the familywise error rate may be preferred in situations where high stakes are attached to the inferences drawn from the hypothesis test.

However, MCPs that control for the familywise error rate are less powerful than those procedures that control for the false discovery rate. Moreover, the power of familywise MCPs drastically decreases as the family of hypotheses increases. Procedures controlling for the false discovery rate are generally able to maintain more stable power as *C* increases. Ultimately, deciding on the appropriate definition of the Type I error rate depends on which research questions are being asked and the purpose of the study (Ryan, 1959). Because consequences tied to research in education are often high stakes, the familywise error rate definition will be used for this study.

Power definitions.

Before continuing on, a brief discussion about power is necessary. As a reminder, power is the probability of correctly rejecting the null hypothesis. In the running example, this corresponds to correctly identifying those schools whose academic achievement was significantly different from the criterion. As with Type I error rate, several definitions of power exist: *any-pair* power, *all-pair* power, and *per-pair* power (Shaffer, 1995). Any-pair power is defined as the probability of correctly rejecting at least one false hypothesis in a set of tests

(Ramsey, 1978). Any-pair power approximates the power of an omnibus *F* test statistic and is most often of interest in exploratory studies (Ramsey, Ramsey, & Barrera, 2010). All-pair power is the probability of correctly detecting all false hypotheses within a family of tests (Ramsey, 1978). It has been recommended that all-pair power is the most appropriate definition of power for confirmatory studies (Ramsey et al., 2010). Per-pair power is defined as the average probability of correctly rejecting a false hypothesis in a family of tests (Einot & Gabriel, 1975).

Properties of MCPs.

There have been dozens of MCPs developed to control for the inflation of the familywise Type I error rate. As stated above, the best MCPs maintain the Type I error rate at α while maintaining the highest power. A MCP is said to be robust if the procedure maintains the Type I error rate at or below α even when the theoretical assumptions of the procedure are violated (Games & Howell, 1976). MCPs that maintain strong Type I error control are to be preferred over those procedures that maintain weak Type I error control. In addition to strong Type I error control, researchers should be concerned with the power of the procedure. There are several characteristics that ensure strong Type I error control while increasing the power of MCP. In the following section, two of these characteristics are discussed: protected vs unprotected MCPs, and simultaneous vs sequential MCPs.

Protected vs. unprotected.

A MCP is said to be *protected* if a significant omnibus test is required before the procedure can be utilized (Seaman et al., 1991). For instance, when testing all pairwise comparisons in a one-way *ANOVA* setting, a significant omnibus *F* test may be necessary first. This omnibus test indicates that at least one pairwise comparison is significant.

If no omnibus test statistic is needed, then the MCP is said to be *unprotected*. In general, protected MCPs are more powerful than unprotected procedures (Seaman et al., 1991). Tukey's HSD test is an example of a protected test while Fisher's Least Significant Difference (LSD) is an example of an unprotected test (Seaman et al., 1991).

Simultaneous vs. sequential.

MCPs that correct for the inflation of the Type I error rate in one step are known as *simultaneous* procedures. Simultaneous MCPs use a single, adjusted α for all hypotheses. Simultaneous procedures tend to be some of the oldest MCPs (Toothaker, 1991). On the other hand, a *sequential* procedure is any procedure that tests two or more stages of a hypothesis or a procedure that depends on a statistic other than the comparison itself (Seaman et al., 1991; Toothaker, 1991). In general, sequential procedures are more powerful than simultaneous procedures (Seaman et al., 1991; Strassburger & Bretz, 2008).

Sequential MCPs may be either step-up or step-down procedures (Brown & Russell, 1997). Step-down procedures begin by comparing the smallest *p*-value

to α' and, assuming rejection of the null hypothesis, iteratively compare each subsequently larger *p*-value to α' until a null hypothesis is retained.

In the next section, two classes of MCPs are discussed: MCPs derived from the Bonferroni inequality and MCPs based upon various range distributions. To be clear, these two classes are not encompassing of all MCPs. For example, methods exist which utilize resampling procedures or graphical analysis. However, these two classes of MCPs have traditionally been the most popular MCPs when multiple hypotheses are tested on a set of means.

Simultaneous Bonferroni based MCPs.

As stated above, when independent hypotheses are tested the familywise Type I error rate inflation may be calculated using Equation 3. Equation 7 adjusts α so that the Type I error rate for a family of tests will not exceed the α selected by the researcher.

$$\alpha' = 1-(1-\alpha)^{1/C}. \qquad (7)$$

The adjusted alpha level is denoted as α'. This is called Šidàk's equation and it controls for the familywise Type I error rate inflation (Šidàk, 1967). To demonstrate, suppose a researcher was testing four hypotheses and wanted to maintain the familywise α at .05. Šidàk's equation would produce an α' of .0127. To test for significance, *p*-values associated with the *t*-tests used to test the four hypotheses would be compared to an α' of .0127, as opposed to the nominal value of .05.

Šidàk's equation maintains strong Type I error control but assumes that all comparisons are independent of one another. This is a result of Šidàk's equation

being derived from Equation 3. A benefit of Šidàk's equation, and all equations derived from it, is that it may be used for any set of multiple *p*-values (Ludbrook, 1998). However, because Šidàk's equation involves the use of a fractional power it failed to gain favor in the pre-computer days (Abdi, 2010). In addition, Šidàk's procedure is a conservative method in that it controls the Type I error rate inflation at a value less than α at the cost of a significant loss of power (Abdi, 2010).

Bonferroni's procedure.

Dunn (1961) popularized a computationally simpler method of controlling the familywise Type I error rate via Bonferroni's inequality (Bonferroni, 1936). This method is alternatively called Boole's inequality or Dunn's approximation (Dunn, 1961). Bonferroni's inequality is the first linear term of the Taylor series expansion of the Šidàk equation (Abdi, 2010). Bonferroni's inequality is written as:

$$\alpha' \approx \frac{\alpha}{C}. \qquad (8)$$

After obtaining *p*-values, Bonferroni's procedure also may be used to directly adjust *p*-values by multiplying the *p*-values by the number of hypotheses. It should be noted that this may result in individual adjusted *p*-values greater than one. This may occur when a large number of hypotheses are being tested and, in such a situation, the adjusted *p*-values should be rounded down to one (Abdi, 2010). As an illustration, suppose the following *p*-values are obtained: .02, .04, and .9. The adjusted Bonferroni adjusted *p*-values would be .01(3) = .03, .04(3) =

.12, and .9(3) = 2.7. The same decision about the null hypothesis will be obtained whether α or the *p*-values are adjusted.

Bonferroni's inequality and Šidàk's equations are related to one another:

$$1-(1-\alpha)^{1/C} \geq \frac{\alpha}{C}. \quad (9)$$

The above inequality states that the adjusted α produced by Šidàk's equation will always be greater than or equal to the adjusted α produced by Bonferroni's inequality. In other words, Šidàk's equation will always be more powerful than Bonferroni's inequality (Abdi, 2010). Empirical evidence, however, suggests the difference in power is very small (Abdi, 2010). As is the case with Šidàk's equation, Bonferroni's inequality maintains strong Type I error control. Similar to Šidàk's equation, Bonferroni's inequality assumes independence of comparison. Unfortunately, in exchange for strong control of the Type I error rate, the Bonferroni MCP sacrifices significant power – particularly as the number of hypothesis tests increases (Gelman et al., 2012; Lu & Westfall, 2009). Other procedures, derived from the Bonferroni inequality, have been developed that produce increased power.

Sequential Bonferroni based MCPs.

Holm's procedure.

Holm's procedure is an example of a sequential step-down procedure for controlling the familywise Type I error rate at or below α (Holm, 1979). To perform Holm's procedure, one obtains the *p*-values from a family of statistical tests. As with Šidàk's and Bonferroni's procedures, these values may be obtained from any test that produces a *p*-value (Holland & Copenhaver, 1988). Holm's

procedure begins by ordering *p*-values obtained from multiple hypothesis tests from smallest to largest. The first *p*-value is then compared to $\frac{\alpha}{C}$. If this *p*-value is larger than $\frac{\alpha}{C}$, then the null hypothesis is retained along with all subsequent null hypotheses and the procedure is terminated. However, if this *p*-value is smaller $\frac{\alpha}{C}$, then the null hypothesis is rejected and the next largest *p*-value is then compared to $\frac{\alpha}{C-1}$. If this hypothesis is rejected, the next largest *p*-value is compared to $\frac{\alpha}{C-2}$. These comparisons continue until a null hypothesis is retained or the smallest *p*-value is compared to α (Holm, 1979).

Similar to Bonferroni's procedure, Holm's procedure can also modify *p*-values directly by multiplying the *p*-value by *C-i+1*, where *i* is an index of the step associated with the *p*-value. For instance, if ten comparisons are being made and one wished to adjust the third smallest *p*-value, the researcher would multiply that *p*-value by 10-3+1. Holm's procedure will always be more powerful than Bonferroni's inequality (Aikin & Gensler, 1996). In addition, Holm's procedure makes no logical assumptions about the hierarchy of the hypotheses to be tested and does not assume independence of comparisons (Seaman et al., 1991). As a result, Holm's procedure may be used whenever a *p*-value is available or as Seaman et al. (1991) stated it may be used in a "virtually limitless variety of inferential statistical contexts" (p. 585).

Holm's procedure does share with Bonferroni's inequality the undesirable attribute of occasionally producing adjusted *p*-values greater than one. As with Bonferroni's inequality, if this occurs, the adjusted *p*-value should be rounded

down to one. The Holm's procedure may be modified to include Šidàk's equation (Abdi, 2010). This is called the Šidàk-Holm's procedure and is slightly more powerful than Holm's procedure and will not produce adjusted *p*-values greater than one.

Hochberg's procedure.

Step-up procedures, on the other hand, compare the largest *p*-value to α' and, upon retention of the null hypothesis, continue iteratively to the next largest *p*-value until a null hypothesis is rejected. Step-up procedures are based off the Simes' inequality (1986) for independent comparisons:

$$p_{(i)} > \frac{i\alpha}{C} = 1 - \alpha, \tag{10}$$

where *C* is the number of comparisons to be made and *i* is an integer between 1 and *C* corresponding to the rank ordered *p*-values. Simes' inequality itself has weak control of the Type I error rate (Levin, 1996). However, the step-up procedures derived from Simes' inequality have demonstrated strong control of Type I error rate (Klockars & Hancock, 1992).

Monte Carlo simulation studies have demonstrated that step-up procedures are empirically more powerful than step-down procedures, particularly when a large number of null hypotheses are false (Dunnett & Tamhane, 1992; Hochberg & Tamhane, 1987; Horn & Dunnett, 2004). The difference in power between step-down and step-up procedures increases with the number of hypotheses to be tested (Dunnett & Tamhane, 1992). Two examples of step-up MCPs are Hochberg's (1988) and Hommel's procedures (1988).

Hochberg's (1988) MCP is a sequential MCP derived from Bonferroni's inequality. To perform Hochberg's procedure, one obtains the *p*-values from a family of statistical tests. Hochberg's approach begins with the ordering of the statistical test's *p*-values from largest to smallest. The largest *p*-value is compared to α. If the first *p*-value is less than α, the null hypothesis is rejected, the procedure is terminated and all remaining hypotheses are rejected. However, if the largest hypothesis is retained, then the second largest *p*-value is evaluated against $\frac{\alpha}{2}$. This process continues iteratively until a hypothesis is rejected or all hypotheses are tested.

Hommel's procedure.

Hommel's (1988) procedure is another sequential that utilizes several logical decision steps, making it a more complex procedure then Hochberg's procedures. Let *C* be the total number of hypotheses, *i* be the number of hypotheses considered at a given step, and *k* index each of the hypotheses by *p*-value beginning with the smallest *p*-value. Beginning with $i = 1$, the following equality is evaluated:

$$p_{C-i+k} > \frac{c\alpha}{i}. \tag{11}$$

If Equation 11 is true, we move next to $i = 2$ and so on until either *j* is equal to *C* or Equation 11 is found to be not be true. Each *p*-value is then compared against:

$$p_i \leq \frac{\alpha}{i}, \tag{12}$$

where *i* is the largest value for which Equation 11 is true.

If no values of *i* exist for which Equation 11 is true than all hypotheses are rejected. (Hommel, 1988; Shaffer, 1995). To illustrate, consider an example given by Nee (2014), where, the *p*-values for three hypotheses are p_1 = .024, p_2 = .03, and p_3 = .073. Beginning with $i = 1$, we evaluate the *p*-value corresponding to $p_{C=3-i=1+k=1}$, which is p_3 or .073. Because this *p*-value is greater than $\frac{C\alpha}{i} = \frac{1(.05)}{1}$ = .05, we continue to $i = 2$. We now evaluate two *p*-values $p_{C=3-i=2+k=1}$ and $p_{C=3-i=2+k=2}$, which correspond to p_2 and p_3. P_2 is compared to $\frac{1(.05)}{2}$ and p_3 to $\frac{2(.05)}{2}$. Again, both *p*-values are greater than their comparison values. When $i = 3$ we find that the inequality in Equation 11 no longer holds, and as a result a value of two is chosen for *i* in Equation 12, giving an adjusted *p*-value of .025. Finally, any *p*-value less than .025 is rejected.

When hypotheses are logically independent, Hommel's procedure will be slightly more powerful than Hochberg's procedure (Shaffer, 1995). Both procedures will always be more powerful than the Bonferroni and Holm's procedures.

Range based MCPs.

MCPs have been developed to specifically control for normally distributed means using several related range distributions: Student's *t*, studentized *q*, and *F* distribution. Test statistics drawn from these distributions are related through the following equality:

$$F = t^2 = \frac{q^2}{2}, \tag{13}$$

where the within groups degrees of freedom for the *F* distribution are equal to the total level-one sample size minus the level-two sample size, the between groups

degrees of freedom for the *F* distribution are equal to the number of level-two units minus one, the degrees of freedom for the Student's *t* distribution is equal to the total level-one sample size minus one, and the degrees of freedom for the *q* distribution are equal to the total level-one sample size minus the level-two sample size.

Tukey's HSD.

The Tukey Honestly Significant Difference (HSD) procedure adjusts for multiplicity by utilizing the studentized range distribution (Toothaker, 1991; Tukey, 1953). The Tukey HSD is a simultaneous MCP that compares a *t*-statistic against a single *q* critical value:

$$\frac{q_{j,df}}{\sqrt{2}}, \tag{14}$$

where *j* is the total number of level-two units being tested and *df* is the within groups degrees of freedom defined above (Toothaker, 1991). The numerator for Equation 14 is drawn from a table of critical values of the studentized range distribution. The null hypothesis is rejected when the absolute value of the *t* statistic exceeds or is equal to this critical value. Because the studentized *q* distribution has thicker tails than the *t* distribution, Tukey's HSD controls for multiplicity by taking advantage of the fact that the *q* distribution necessitates larger evidence to declare a significant mean difference. While Tukey's HSD was designed to control for multiplicity in the all pairwise comparisons scenario, this MCP can easily be adopted to comparing several means against a criterion by obtaining t-values drawn from single-sample *t*-tests.

Scheffé's Procedure.

The Scheffé MCP (1959) is a simultaneous MCP based upon the *F* distribution. The Scheffé MCP compares *t* values drawn from any linear combination of means against the critical value:

$$\sqrt{(j-1)F_{j-1,df}}, \tag{15}$$

where *F* is drawn from a table of critical values for the *F* distribution, and the remaining terms are defined above.

Tukey's HSD is more powerful than Bonferroni's MCP and Scheffé's MCP when making all pairwise comparisons. Scheffé's MCP gains power as the number of hypotheses increase and is more powerful than Tukey's HSD when level-two units have unequal sample sizes (Shaffer, 1995). Both range procedures may be used as a protected or unprotected MCP. Both Tukey's HSD and Scheffé's MCP tend to be slightly less powerful than the sequential derivations of the Bonferroni procedure.

Factors affecting MCPs.

A variety of factors may affect the Type I error control or, more commonly, the power of MCPs. These factors may arise from the MCP's assumptions, an underlying statistical test, or as the result of properties of the sample and the decisions made during the construction of the research design. In the following section, these factors are categorized as either statistical assumptions or practical considerations.

Assumptions.

MCPs either have their own assumptions about the data, as is the case with the range procedures, or are affected by the assumptions underlying the statistical tests from which the *p*-values are obtained, as is the case with the Bonferroni based procedures. When making multiple comparisons among means, the assumptions of both classes of MCPs are similar because the *p*-values for the Bonferroni based procedures are often taken from *t*-tests, which share many of the assumptions of the range MCPs. Specifically, these procedures assume independence of observations and normally distributed data. The independent samples *t*-test, which is appropriate when making pairwise comparisons, contains and additional assumption regarding variance homogeneity among the level-two units. It should be noted that the fixed effects regression approach to testing hypotheses about level-two units makes the same assumptions. Research has generally shown that the MCPs described above are not greatly affected by violations to the normality assumption (Brown & Russell, 1997; Einot & Gabriel, 1975; Ramsey et al., 2010).

Violating the variance homogeneity assumption has drastic effects on the performance of MCPs (Nashimoto & Wright, 2008). While the MCPs described thus far are generally able to maintain the Type I error rate at or below α when this assumption is not met, the power to detect true pairwise mean differences is severely reduced when even moderate heterogeneity of variance is present (Games & Howell, 1976; Hsiung & Olejnik, 1994; Kromrey & La Rocca, 1995). When comparing every level-two unit mean to a criterion, those level-two units

with larger within-groups variances will have less power to detect true differences from the criterion as compared to those level-two units with smaller variances, even when the mean difference from the criterion are identical. This in turn may affect the all-pairs power of an MCP to detect true differences from the criterion for the family of tests under consideration. However, this has not been evaluated empirically.

Practical considerations.

A number of practical considerations may also affect the performance of MCPs. These factors may be the result of the research design, available resources, or the limitations of the sample itself. Factors that may affect the performance of MCPs include the total sample size across all level-one units, the number of level-two units, and the definition of power used.

Level-one sample size.

Increasing the total level-one sample size across all level-two units will increase the power of all MCPs under consideration. This has been consistently demonstrated in the literature (Hsiung & Olejnik, 1994; Kromrey & La Rocca, 1995; Olejnik, Li, Supattathum, & Huberty, 1997; Ramsey, 1981; Ramsey et al., 2010; Seaman et al., 1991).

Unequal level-one sample sizes.

When the level-one sample sizes vary between the level-two units, the power of MCPs will be adversely affected. Those level-two units with smaller level-one sample sizes will have less power to reject the null hypothesis as

compared to those level-two units with larger sample sizes, even when the level-two means are equivalent. This in turn reduces the all-pair power of MCPs.

Level-two sample size.

Increasing the number of level-two units drastically decreases the power of the MCPs described above (Kromrey & La Rocca, 1995; Olejnik et al., 1997; Seaman et al., 1991). This can be seen by revisiting Equation 3. To begin with, assume six hypotheses are evaluated. To illustrate, using Bonferroni's procedure each hypothesis would be evaluated against an α' of .05/6 = .0083. If the number of hypotheses is increased to 45 each hypothesis would be tested against an α' of .05/45 = .0011. As can be seen, the inverse relationship between the number of level-two units and α' substantially reduces the probability of correctly rejecting the null hypothesis. The critical values for the range based procedures are based, in part, by the number of level-two units. A greater number of level-two units results in a higher critical value, decreasing the power to detect true differences. This problem seriously limits the applicability of MCPs in educational research situations when a large number of hypotheses are tested.

Power.

As stated above, there are three popular definitions of power (all-pair, any-pair, and per pair) and the choice of which definition to use will affect the performance of an MCP. Typically, MCPs demonstrate stronger any-pair than all-pair power, and usually by a large amount (Kromrey & La Rocca, 1995; Olejnik et al., 1997), with this discrepancy increasing as a direct function of the number of hypotheses to be tested (Horn & Dunnett, 2004). Per-pair power

depends on the number of hypotheses to be tested but generally falls between any-pair and all-pair power. Increasing the number of hypotheses to be tested increases the per-pair power (Horn & Dunnett, 2004). Analogous to choosing which definition of the Type I error rate to use, selecting which definition of power to use depends on the purpose of the research questions being asked. As a result, the chosen definition of power may be made independently of the definition of the Type I error rate.

Among the multiple factors that may influence the performance of MCPs, two factors, variance heterogeneity and the number of level-two units, are of particular concern. These factors reduce the power of MCPs so substantially that they render MCPs, for all intents and purposes, useless in detecting true mean differences. A method outside traditional MCPs is needed that maintains strong control of the Type I error rate while preserving power when variance heterogeneity and a large number of level-two units are present. Up to this point, all level-two parameter estimates have been treated as fixed parameters. One possible solution may be to treat the level-two units as *random* effects through the use of multilevel models (MLMs).

Multilevel Models

Overview.

MLMs expand on fixed effects models by allowing parameters to vary among higher-level units. The MLM corresponding to Equation 1, where we wish to account for the effect of a grouping variable, may be expressed as a hierarchical linear model (Raudenbush & Bryk, 2002):

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad (16)$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j}, \quad (17)$$

where y_{ij} is the achievement for student *i* for school *j*, β_{0j} is the average achievement for the *j*th school, and ε_{ij} is the level-one error term distributed as $\underline{N}(0,\sigma^2)$. For the level-two parameters, γ_{00} is the average population school academic achievement and μ_{0j} is the deviation from the grand mean for school *j* and is distributed as $N(0,\tau_{00})$.

Intraclass correlation coefficient.

The intraclass correlation coefficient (*ICC*) is a measure of the extent to which the variability in the outcome is due to variability in the higher-level units (Raudenbush & Bryk, 2002). The ICC may be thought of as a measure of the necessity of modeling data in a multilevel format. The *ICC* statistic ranges from zero to one. Research has indicated that *ICC* values between .15 and .3 are most often observed in educational research (Hedges & Hedberg, 2007a; Hedges & Hedberg, 2007b). The *ICC* is expressed as:

$$ICC = \frac{\tau_{00}}{\tau_{00}+\sigma^2}. \quad (18)$$

A value of zero indicates that none of the variance in the lower level was due to variability in the higher-level units. Using the running example, an *ICC* of zero would indicate that any variability in student achievement was due to within student variability and the school had no effect on achievement. This effect is analogous to aggregating over the effect of schools. When the *ICC* is zero or it is not necessary to model the data hierarchically.

An *ICC* of one, on the other hand, indicates that variability in the outcome can be completely explained by higher-level membership. This corresponds to the scenario in which any variability in academic achievement can be totally explained by the school a student attends.

Random effects.

In recent years, MLMs have become more popular in educational research. Federal educational initiatives such as the *No Child Left Behind Act of 2001* and the *Elementary and Secondary Education Act* have inspired researchers to investigate how student level data, such as standardized test scores, may be used to infer knowledge on the performance of schools and teachers (Shaw, 2012). The line of research, spurred on by these government policies, acknowledges the inherent nesting of students within schools or teachers. Thus, researchers have increasingly focused on developing models that isolate the teacher or school effects on student performance above and beyond what may be explained by student characteristics such as maturation or demographic information. As a result, MLMs present themselves as an attractive option for applied researchers and, increasingly, researchers have parameterized the effects of schools and teachers as random. Two prominent examples of this line of research are value added models and models for the accreditation of various institutions (Goldstein & Thomas, 1996; Ohlessen et al., 2007; Shaw, 2012)

MLMs are appealing to researchers for several reasons. First, MLMs represent a compromise between aggregating over level-two units and the fixed effects model seen in Equation 1 (Gelman & Hill, 2007; Gelman et al., 2012;

Ohlessen et al., 2007; Raudenbush & Bryk, 2002). Aggregating over the levels of a categorical variable assumes that the level-two units are homogenous in their effect on the outcome (Pedhazur, 1997). In the running example these fixed effects models would assume that all schools had an equivalent effect on student achievement. Treating the effects of schools as fixed assumes that the effect of each school is completely independent of every other school. For example, although schools in a region may share characteristics that affect student achievement the influence of these commonalities would be ignored in a fixed effects model. MLMs acknowledge that individual level-two units provide information about the effects of other level-two units and incorporate this knowledge into the parameter estimates by shifting the level-two unit mean estimates towards the grand mean (Gelman et al., 2012).

Second, because the level-two units are assumed to be sampled from a larger distribution of level-two units, MLMs allow researchers to generalize the results to level-two units not considered in the analysis. Fixed effects analysis, on the other hand, limits generalizations to only the level-two units included in the study.

Third, MLMs allow the inclusion of predictors at multiple levels. For example, in a two-level scenario the gender or SES of the students might be examples of a level-one predictor and the region in which the school was located or whether the school was public or private might be examples of a level-two predictor. Neither aggregate nor fixed effects models can incorporate level-one and level-two predictors simultaneously.

Lastly, when a large number of higher-level units are available, there may not be sufficient degrees of freedom for a fixed effects model to arrive at a viable solution (Tabachnick & Fidell, 2013). To understand this, recall that for ordinary least squares linear regression the residual degrees of freedom are equal to $n-(k+1)$, where k is the number of independent variables. Essentially, residual degrees of freedom indicate the amount of information remaining to estimate variability in the dependent variable. If the residual degrees of freedom is zero or negative, the variability of the dependent variable cannot be estimated. As an example, suppose a researcher wishes to predict student academic achievement using one of ten schools and one of five levels of funding as separate, categorical independent variables. A sample size of at least 16 students would be necessary to run this model otherwise the residual degrees of freedom would be zero or negative and the model could not be estimated. The minimum sample size increases as the number of higher-level units increases. MLMs avoid this problem and conserve the residual degrees of freedom spent on the level-two units by treating the level-two units as a single variance parameter to be estimated (Raudenbush & Bryk, 2002).

Hypotheses about level-two means.

Random effects are typically included as a nuisance parameter to account for unexplained variance rather than as a parameter worth exploring in and of itself. However, researchers have used estimates of random level-two means to make inferences about level-two units. For example, Raudenbush and Willms

(1995) investigated using multilevel models to estimate school effects. Goldstein and Thomas (1995) conducted a similar study on schools in the United Kingdom.

In order to calculate parameter estimates for the means of the level-two units, one may follow the procedure detailed by Raudenbush and Bryk (2002). First, letting n_j represent the level-one sample size for level-two unit j, we define the sample mean for each level-two unit as:

$$\bar{y}_{.j} = \beta_{0j} + \bar{\varepsilon}_{.j}, \tag{19}$$

where

$$\bar{\varepsilon}_{.j} = \sum_{i=1}^{n_j} \varepsilon_{ij} / n_j \ . \tag{20}$$

The variance for Equation 20, referred to as the error variance, is defined as:

$$V_j = \sigma^2 / n_j, \tag{21}$$

where σ^2 is the level-one error variance. The variance for Equation 19 can then be defined as:

$$\Delta_j = \tau_{00} + V_j . \tag{22}$$

Assuming that all Δ_j are known and equal level-one sample sizes between the level-two units, the estimated grand mean is defined as:

$$\hat{\gamma}_{00} = \sum \Delta_j^{-1} \bar{y}_{.j} / \sum \Delta_j^{-1}, \tag{23}$$

with variance:

$$\text{var}(\hat{\gamma}_{00}) = (\sum \Delta_j^{-1})^{-1} . \tag{24}$$

We can then calculate an estimate for each level-two mean by:

$$\beta_{0j}^{*} = \lambda_j \bar{y}_{.j} + (1 - \lambda_j) \hat{\gamma}_{00}, \tag{25}$$

where λ_j is the reliability of the estimated sample mean for level-two unit *j* and defined as:

$$\lambda_j = \text{var}(\beta_{0j}) / \text{var}(\bar{y}_{.j}) = \tau_{00} / (\tau_{00} + V_j) . \quad (26)$$

We can then construct a confidence interval for β^*_{0j} by first defining its variance:

$$V^*_j = (V^{-1}_j + \tau^{-1}_{00})^{-1} + (1 - \lambda_j)^2 \text{var}(\hat{\gamma}_{00}) . \quad (27)$$

The 95% confidence interval for β_{0j} (assuming normality) is then:

$$95\%CI(\beta_{0j}) = \beta^*_{0j} \pm 1.96 V^{*1/2}_j . \quad (28)$$

The above equations have assumed that the variance parameters σ^2 and τ_{00} are known *a priori*, which is almost never the case in practice. More commonly, the maximum likelihood of variance parameters is estimated using iterative methods, such as the expectation-maximization algorithm (Raudenbush & Bryk, 2002). Maximum likelihood estimators select estimates of σ^2 and τ_{00} that maximize the likelihood of the observed sample data.

Using Equation 28, researchers may then test the hypothesis that any level-two mean differs significantly from a criterion by inspecting whether the confidence interval for the level-two unit encapsulates the criterion. In the ongoing example, this is done by evaluating whether zero is contained in the confidence interval for β_{0j}. If the confidence interval excludes the criterion, there is evidence that the level-two mean differs significantly from the criterion. Alternatively, researchers could test the hypothesis that a level-two mean is different from a criterion by deriving a Wald *z* statistic and testing a point hypothesis about the level-two mean.

There is an argument to be made about the appropriateness of declaring as random a level-two mean that was previously treated as a fixed effect. The above formulas demonstrate that the estimate of a mean will often differ depending on whether it is treated as a fixed or random effect. Arguments regarding the correctness of treating an effect as fixed or random outside of theoretical grounds are beyond the scope of this paper. Having said that, this paper largely avoids this issue because the focus is on the probability of making a correct decision rather than providing the most accurate estimates of the level-two means. Further, as discussed below, differentiating between fixed and random effects is resolved when variables are viewed through the Bayesian lens because it treats all variables as random.

Three level models.

Although this paper focuses on testing hypotheses about the level-two means, it should be noted that it is straightforward to expand this model to a three level hierarchical model. Building upon the ongoing example, suppose each of *i* students are nested in one of *j* schools which, in turn, are nested within one of *k* districts. The three-level model has the added benefit of accounting for variance between level-two units that are due to a shared level-three unit. The three level MLM may be expressed as:

$$y_{ijk} = \beta_{0jk} + e_{ijk} \tag{29}$$

$$\beta_{0jk} = \delta_{00k} + U_{0jk} \tag{30}$$

$$\delta_{00k} = \gamma_{000} + V_{00k}, \tag{31}$$

where y_{ijk} is the academic achievement of student *i* attending school *j* nested within district *k*, e_{ijk} is the level-one error term distributed as $N(0,\sigma^2)$, β_{0jk} is the school level deviation from its district level mean, δ_{00k}, U_{0jk} is the level-two error term distributed as $N(0, \tau^2_{U0})$, γ_{000} is the grand mean, and V_{00k} is the level-three error term distributed as $N(0, \tau^2_{V00})$. Testing hypotheses about the means of the higher-level units follows from the procedures detailed above in Equations 19 through 28 (Raudenbush & Bryk, 2002).

Calculating the *ICC* for three level models is a bit more complicated than calculating the *ICC* for two level models. This is because two separate *ICCs* must be computed. The first, referred to as the *ICC2*, is a measure of the total variation in the outcome that is due to the level-two unit (Hoffman, 2016). Under the running example, the *ICC2* corresponds to the variation in academic achievement that is attributable to school membership. Previous research has indicated that *ICC2* values between .15 and .25 are observed in educational data (Hedges & Hedberg, 2007A). The *ICC2* is expressed as:

$$ICC2 = \frac{\tau^2_{V00}+\tau^2_{U0}}{\tau^2_{V00}+\tau^2_{U0}+\sigma^2}. \quad (32)$$

The second, referred to as the *ICC3*, is a measure of the total variation in level-two that is due to level-three membership (Hoffman, 2015). Using the ongoing in example, the *ICC3* is a measure of the variability in academic achievement between schools that is due to the district a school belongs. The *ICC3* is expressed as:

$$ICC3 = \frac{\tau^2_{V00}}{\tau^2_{V00}+\tau^2_{U0}}. \quad (33)$$

There is a lack of research on common values of the *ICC3* in the social sciences however it has been suggested that the *ICC3* is typically lower than the *ICC2* (Siddiqui, Hedker, Flay, & Hu, 1996).

More complex models, which may include predictors at both levels, additional random effects, and cross-classified levels of nesting, are available and, if properly specified, improve the parameter estimates of MLMs (Gelman et al., 2012). For example, value added models typically nest students both within teachers and time points, in addition to including a variety of additional covariates.

Using MLMs as a MCP.

Several authors (Gelman et al., 2012; Kruschke, 2011; Raudenbush, 1988) have speculated that MLMs may provide some inherent control for Type I error rate inflation by shifting the estimates of the level-two means towards the grand mean, a phenomenon termed *shrinkage*. Shrinkage makes it more difficult to declare any a level-two unit as significantly different from the grand mean (Gelman et al., 2012; Raudenbush & Bryk, 2002). Suppose that no level-two mean differs significantly from the criterion of zero in our example, which is the true population mean. In this situation, each level-two mean estimate would be pulled towards the criterion, decreasing the probability of committing a Type I error. On the other hand, suppose each level-two mean in our sample was ten points higher than the true, population mean of zero. In this scenario, the estimated level-two means would be shrunk towards the estimated population mean of ten, and the confidence intervals surrounding the level-two mean

estimates would exclude the true population mean of zero. These two examples encompass the extreme situations in which all level-two means are either equal or different than the true population mean. For scenarios in between, where only a portion of the level-two means are different from the criterion, the non-zero level-two means would not be drawn as close to the criterion as the level-two units with means of zero.

In fact, shrinkage is similar to how MCPs operate. For example, assume there are three level-two units, all with a standard error equal to 1. The null hypothesis that a level-two mean is significantly different than the aggregate mean, H_0: $\mu_j = 0$, is rejected if $\left|\frac{\bar{x}_j}{se}\right| > t_{(\frac{\alpha}{2},df)}$. If alpha is equal to .05 and the sample size for each level-two unit is 30, H_0 is rejected if $\left|\frac{\bar{x}_j}{se}\right| > 2.045$. Applying the Bonferroni correction, the mean difference necessary to reject H_0 increases to $\left|\frac{\bar{x}_j}{se}\right| > 2.541$. With shrinkage, the mean estimates are shrunk towards the aggregate mean and the amount of shrinkage for a given level-two unit may be calculated as:

$$\bar{x}_{adj} = \frac{\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau_{00}}\bar{x}_{agg}}{\frac{n}{\sigma^2} + \frac{1}{\tau_{00}}}, \tag{34}$$

where $\bar{x}_{adj}$ is the shrinkage adjusted mean and $\bar{x}_{agg}$ is the aggregate mean (Gelman & Hill, 2007). Using the above value for the standard error, suppose $\bar{x}_1$ is equal to 3.5, σ^2 is equal to 30, τ_{00} is equal to .5, and $\bar{x}_{agg}$ is equal to 1. With shrinkage, the *t* critical value remains constant at 2.045. The fixed effect test of

H_0: $\mu_1 = 0$ would be rejected in a fixed effects model because $\left|\frac{3.5}{1}\right| > 2.045$. However, shrinkage would alter the estimates of $\bar{x}_1$ to 1.833. As a result, with a random effects model H_0 would be retained since $\left|\frac{1.833}{1}\right| < 2.045$. As can be seen, shrinkage decreases the probability of committing a Type I error.

Using MLMs as a MCP may also address some of the issues that follow with testing a large number of hypotheses. MLMs parameter estimates become more accurate as the number of level-two units increase (Baldwin & Fellingham, 2013; Raudenbush & Bryk, 2002; McNeish & Stapleton, 2016; Stegmueller, 2013). Further, as seen in Equation 34, the degree of shrinkage for a given level-two mean is not directly influenced by the level-two sample size, rather it is solely a function of the within and between level variances. This suggests that the power of using MLMs as an MCP would not be adversely effected by increasing the number of tested hypotheses.

Unfortunately, there has been a lack of research on how effectively frequentist MLMs operate as a control for multiplicity. As noted above, this is likely because researchers have not been interested in individual level-two parameter estimates beyond the extent to which the random effect controls for unexplained variance. Additionally, it may not seem intuitive to make an inference about level-two means because the level-two units are assumed to be randomly sampled from a larger population. However, because confidence intervals are used to test hypotheses about level-two means, testing several level-two means against a criterion should result in some Type I error rate inflation above and beyond the protection provided by the shrinkage phenomenon.

Few studies have investigated the error rates of making classification decisions based on using MLMs. Because the variance components of frequentist MLM are commonly estimated via maximum likelihood, it is not uncommon to obtain estimates of the variance components that are equal to zero (Bayarri & Berger, 2004); which would make the evaluation of the Type I error difficult. One study, by Schochet and Chiang (2010), examined the error rate of complex MLMs in which students were nested within teachers. The models examined had several predictors at both the teacher and student level in addition to time varying random effects. The authors found that these model exhibited Type I error rates greater than α.

Additionally, MLMs provide no adjustment for Type I error inflation when τ_{00} is equal to zero, which is the situation in which protection against multiplicity would be most desired. This can in Equation 34 by setting τ_{00} to zero, which is equivalent to aggregating across the level-two units.

Bayesian Paradigm

An alternative approach to controlling the multiplicity problem may be to adopt a Bayesian perspective to statistical testing. The key difference between the frequentist and Bayesian paradigms is that the Bayesian paradigm allows researchers to treat all parameters as random variables arising from some distribution while the frequentist paradigm assumes that parameters have a single, fixed value (Kaplan, 2014; Lynch, 2007). Additionally, the Bayesian paradigm allows researchers to assign prior distributions to all parameters in a given model. Returning to Equation 2, rather than treating β_0 as a parameter with a single, fixed

value, the Bayesian model would assign a prior distribution to β_0 that encompasses all possible values of β_0. Further, the Bayesian approach to statistics allows the researcher to assign prior distributions to the parameters of prior distributions, called hyperparameters (Kaplan, 2014). For instance, the frequentist linear regression model assumes that there is an error term associated with every model and that is distributed $N(0,\sigma^2)$; this itself can be conceptualized as a prior distribution. With Bayesian analysis, it is possible to assign prior distributions to the mean and variance of the distribution of the error term. Note that this allows hyperparameters to have their own hyperparameters. It is not necessary to specify a prior distribution for all hyperparameters; hyperparameters may alternatively be assigned fixed values (Kaplan, 2014).

To understand the rationale to Bayesian statistics, let Y be a random variable that takes on the observed data, y, and let θ be unknown to the researchers and represent a parameter or set of parameters that define a probability model meant to explain the observations, y (Kaplan, 2014). The likelihood of the parameter estimates given the data may be written as $L(\theta|y)$. Further, the probability of obtaining y given θ, $p(y|\theta)$, is proportional to the likelihood and is known as the posterior distribution of θ (Kaplan, 2014). In the Bayesian framework, θ is assumed to be random and have its own probability distribution (Kaplan, 2014). The Bayesian approach to statistics also makes the assumption of exchangeability which "implies that the subscripts of the vector of data (e.g., $y_1, y_2, \ldots, y_n$) do not carry information that is relevant to describing the probability distribution of the data" (Kaplan, 2014, p. 16). The goal of Bayesian

statistics is to determine the probability distribution that best estimates θ given the data and then to summarize that distribution (Lynch, 2007). Many of these summaries are integrals of the posterior distribution, such as the mean, mode, and variance (Lynch, 2007).

Because both θ and y are assumed to be random, it is possible to model the joint probability of θ and y, $p(\theta, y)$, as the product of $p(y|\theta)$ and the prior distribution of θ, $p(\theta)$. The prior distribution is an acknowledgement by the researcher of what is known or unknown about the parameter of interest. Equation 35 demonstrates this relationship:

$$p(\theta, y) = p(y|\theta)p(\theta). \quad (35)$$

Using Bayes theorem, the posterior distribution of θ can be written as:

$$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}, \quad (36)$$

where $p(y)$ ensures that the posterior distribution integrates to one. Because the purpose of $p(y)$ is to scale the posterior distribution to for a proper density and does not contain model parameters Bayes theorem is often written:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \quad (37)$$

When $p(y|\theta)$ is expressed in terms of an unknown θ for fixed values of y, this term becomes $L(\theta|y)$ and Equation 37 is rewritten as:

$$p(\theta \mid y) \propto L(\theta \mid y)p(\theta). \quad (38)$$

Conceptually, Equation 38 shows that the posterior distribution is the product of the likelihood function and the prior distribution (i.e., *Posterior* $\propto$ *Likelihood x Prior*). Essentially, the values obtained from the likelihood function updates the range of reasonable values that might contain θ (Kruschke, 2013). As

Lynch (2007, p. 50) succinctly writes "the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty." Kruschke (2013, p. 574) describes Bayesian analysis as a process that "reallocates belief toward the parameter values that are consistent with the data and away from the parameter values that are inconsistent with the data."

The Bayesian approach to statistics has a number of advantages compared to the frequentist approach to statistics. For example, because Bayesian analyses seek to summarize the posterior distribution there is no need to rely on asymptotic assumptions to ensure normality, as is common in many frequentist hypothesis tests (Lynch, 2007). Moreover, compared to the frequentist approach to statistics, Bayesian analysis allows for more measures of model fit and provides more information about parameter estimates (Lynch, 2007).

However, the Bayesian paradigm is not without criticism. For instance, the choice of a prior distribution introduces subjectivity into the analysis and the assumption of whether parameters are randomly distributed is questionable (Lynch, 2007). Second, some have argued that there are situations in which parameters do have a true single value, and as a result it is not necessary to place a probability distribution on these parameters. Further, compared to frequentist analysis Bayesian analysis often requires increased computational power (Nashimoto & Wright, 2008).

Elaborating on the first criticism, because researchers have the freedom to choose prior distributions, the choice of prior distributions may unduly introduce the researcher's bias into the analysis that may affect the interpretation of the results (Lynch, 2007). Additionally, two (or more) researchers may select different prior distributions for the same model and data. Because the estimation of the posterior distribution will vary depending on which prior distribution is chosen, the competing researchers may come to different conclusions as a result of their choice of prior distributions (Lynch, 2007).

Bayesian proponents have responded to this criticism in several ways. First, Bayesians argue that all approaches to statistics are subjective to some extent. The choice of α, or the selection of a likelihood function used for a given analysis is subjective as well. Lynch (2007) gives the example that when faced with ordinal data, researchers have the option of choosing a normal likelihood function or a binomial likelihood with a link function. The researcher's choice of function may be due in some part, to the researcher's preferences for the chosen model. Second, Bayesian researchers argue that this uncertainty grants added benefits to Bayesian analysis, namely that the prior distribution can incorporate the findings of previous research (Lynch, 2007). Finally, priors tend to be dominated by the data, particularly when noninformative priors are chosen. As a result, the effect of a prior distribution on the interpretation of the analysis is typically small (Lynch, 2007).

Recall that the second criticism states that because some (or perhaps all) parameters have a fixed value in truth then it is not appropriate to place a

probability distribution on the parameters (Lynch, 2007). For example, suppose the parameter of interest is the average height of a particular classroom of students at a specific point in time; in such a situation it is difficult to argue that true, population average height is anything other than a single value (Lynch, 2007). Bayesian researchers counter this argument by stating that the Bayesian philosophy of statistics is a subjective approach to uncertainty. That is, it does not matter whether or not a parameter is fixed as we are unclear about its true value. As a result, it is irrelevant if a parameter has a true, fixed value (Lynch, 2007).

The final criticism that Bayesian models require greater computational power than their frequentist counterpoints because often it is necessary to use complex sampling methods to estimate the posterior distribution. However, computational power has continued to grow and the Bayesian approach has increased in popularity in concert with this growth (Nashimoto & Wright, 2008).

Prior distributions.

The prior distribution of θ is assigned by the researcher and may be informative or noninformative (Kaplan, 2014). Informative prior distributions assume the researcher has some previous knowledge about the distribution of θ. This information may be drawn from previous research or knowledge about the domain of θ (Kaplan, 2014). For instance, if academic achievement can range from 0 to 100 and it is expected that the majority of students will score in the middle of the distribution, it seems reasonable to specify a prior distribution $N(\mu, \sigma^2)$. Further, it makes little sense to allow the range of the prior distribution to produce negative estimates of academic achievement or estimates of academic

achievement greater than 100. A sensible prior distribution in this example might be $N(50, 10)$.

Informative prior distributions may be further classified as conjugate or non-conjugate priors. Multiplying a conjugate prior distribution by the likelihood results in a posterior distribution from the same family of distributions as the prior (Kaplan, 2014). In contrast, a non-conjugate prior distribution produces a posterior distribution from an unknown family, and it may be difficult, time consuming, or even impossible to derive such distributions. As a result, conjugacy is a desirable property when choosing a prior distribution (Kaplan, 2014).

Noninformative prior distributions, equivalently called objective, vague or diffuse priors, are used when the researcher has little previous knowledge of θ (Kaplan, 2014). As an example, suppose the researcher has no idea whether school J differs significantly from the criterion in academic achievement. A possible non-informative prior distribution may be a uniform distribution, $U[A, B]$, where the hyperparameters A and B are given values of 0 and 100 to account for the minimum and maximum academic achievement scores. When a noninformative prior distribution is used, the posterior distribution will be more heavily influence by the likelihood function and produce estimates that align with those obtained from frequentist inferences. Additionally, noninformative prior distributions are generally non-conjugate.

MCMC sampling.

In practice, it can be impossible to integrate some posterior distributions. This can occur when specifying models containing many parameters or when noninformative and non-conjugate prior distributions are used (Kaplan, 2014; Lynch, 2007). Traditionally, difficulty of integrating the posterior distribution has been one of the factors that has limited the widespread use of Bayesian analysis (Kaplan, 2014). The development of modern sampling methods and increase in computing power have provided researchers an avenue for summarizing posterior distributions (Lynch, 2007). These methods generate a sample from the posterior distribution of interest and summarize these samples to approximate the corresponding integrals (Lynch, 2007). For instance, the expectation of a posterior distribution may be estimated as:

$$E[p(\theta \mid y)] \approx \frac{1}{T}\sum_{i=1}^{T} p(\theta_t \mid y)\,, \tag{39}$$

where T samples of θ are taken from the posterior distribution. For independent and increasing T the approximations of the posterior distribution becomes more accurate (Kaplan, 2014).

Markov chain Monte Carlo (MCMC) sampling algorithms are a set of procedures for summarizing the posterior distribution. MCMCs operate by sampling the domain of all elements with a non-zero probability density for one or more dimensions of a posterior distribution (Kaplan, 2014; Lynch, 2007). The selection of random samples from a distribution is called Monte Carlo integration, whereas Markov chains are the tools used to sample a new value from the posterior distribution (Kaplan, 2014; Lynch, 2007).

A Markov chain is a series of sequential, dependent, random variables in which the conditional probability of drawing a particular variable in the sequence depends only on the immediate previous variable (Kaplan, 2014). Markov chains have the desirable property that, given enough iterations, Markov chains will "forget" the initial distribution from which variables were drawn and converge towards the posterior distribution. Additionally, Markov chains allow the relaxation of the independence assumption of Monte Carlo integration (Kaplan, 2014). Two of the more commonly used algorithms for constructing a Markov chain are the Metropolis-Hastings (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) and Gibbs approaches.

The Metropolis-Hastings algorithm begins by specifying starting values, *s*, for the parameter $\theta^{s=0}$. There are several methods for choosing starting values such as the using the maximum likelihood estimate of the parameter or by using a random number generator to select a starting value. Next, draw a candidate parameter, θ^{c}, from a proposal distribution ($s = 1$). The proposal distribution is chosen by the researcher and commonly selected proposal distributions include the normal or uniform distributions. The next step is to compute the ratio:

$$R = \frac{p(\theta^{c})\,p(\theta^{s-1} \mid \theta^{c})}{p(\theta^{s-1})\,p(\theta^{c} \mid \theta^{s-1})}). \tag{40}$$

R is then compared with a randomly drawn value from a $U[0, 1]$. If *R* is greater than this value, the candidate is accepted as a value draw and θ^{s+1} is set to equal θ^{c}. Otherwise, this candidate is discarded, the previous value of θ^{s} is kept, and another candidate value is drawn (Lynch, 2007). This process is repeated until enough draws have been accepted to converge to the posterior distribution

(Lynch, 2007). The Metropolis-Hastings algorithm has the desirable property of converging towards the posterior distribution regardless of the starting values and the chosen proposal distribution (Kaplan, 2014, Lynch, 2007).

To explain the Gibbs sampler, let $\boldsymbol{\theta}$ represent a vector of parameters, $\{\theta_1, \theta_2, \ldots, \theta_q\}$, where information for $\boldsymbol{\theta}$ is drawn from the prior distribution (Kaplan, 2014). Starting values s are then assigned to these parameters such that $\boldsymbol{\theta}^{(0)} = (\theta_1^0, \theta_2^0, \ldots, \theta_q^0)$. Setting s to $s+1$, the Gibbs sampler then generates a value for $\theta_q^{(s)}$ by sampling from the conditional distribution of that parameter given the current value of all other parameters (Lynch, 2007). This process is described in the following algorithm:

$$
\begin{aligned}
&1. \quad \theta_1^s \sim p(\theta_1 \mid \theta_2^{s-1}, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, y) \\
&2. \quad \theta_2^s \sim p(\theta_2 \mid \theta_1^s, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, y) \\
&\vdots \qquad \vdots \\
&q. \quad \theta_q^s \sim p(\theta_q \mid \theta_1^s, \theta_2^s, \ldots, \theta_{q-1}^s, y)
\end{aligned}
\tag{41}
$$

$q + 1$. Return to step 1.

In words, a value for θ_1 is drawn from the conditional distribution of θ_1 given the data and all other parameters are at start values of 0 (Kaplan, 2014). Next, θ_2 is drawn from the conditional distribution of θ_2 given the data, the current value for θ_1, and the remaining parameters at their start values. The remaining parameters are estimated using the current values for the previous parameters. An iteration is finished upon estimating the last parameter in $\boldsymbol{\theta}$. This process is then repeated until convergence towards the posterior distribution is reached. The process of assessing convergence is discussed in more detail below.

There are benefits and drawbacks to both the Metropolis-Hastings and Gibbs algorithms. Both will converge to the posterior distribution regardless of the chosen starting values (Lynch, 2007). However, it should be noted that "poor" starting values may lead to computational inefficiency. The Metropolis-Hastings algorithm may be more generally applied than the Gibbs sampler because if there is difficulty in deriving the conditional distribution for all parameters in $\boldsymbol{\theta}$ the Gibbs sampler cannot be used. Fortunately, this limitation does not apply to analyzing hierarchical parameters, such as those found in multilevel models. The default MCMC procedure for many Bayesian software programs, such as JAGS (Just Another Gibbs Sampler), WinBugs and the MCMCPack (Martin, Quinn, & Park, 2011), BEST (Krushke & Meredith, 2015), and LaplacesDemon (Statisticat, 2016) packages in *R*, is the Gibbs sampler (Lynch, 2007). As a result, the Gibbs algorithm enjoys more widespread use when estimating multilevel models than the Metropolis-Hastings algorithm.

Ensuring that enough iterations have been performed so that the MCMC algorithm converge to the posterior distributions is an important aspect of Bayesian analysis. An algorithm that has not converged to the posterior distribution may produce inaccurate parameter estimates. Although there are several methods available to assess the convergence of MCMCs, there is no consensus on which is the single best method or criterion for declaring that an algorithm has converged (Kaplan, 2014). Popular convergence diagnostics include trace plots, the Gelman-Rubin convergence diagnostic, and the auto correlation function plot (Kaplan, 2014).

Trace plots are graphical displays of accepted values at each iteration, with the sample value on the ordinate and the iteration on the abscissa. Convergence towards the posterior is demonstrated by a stationary, horizontal line across the plot (Lynch, 2007). The Scale reduction factor or Gelman-Rubin convergence diagnostic (denoted $\hat{R}$) is a test statistic that calculates the ratio of variance in a parameter's Markov chain that is due to within chain variability (Gelman, 1996; Lynch, 2007). Values of $\hat{R}$ close to 1 provide evidence that convergence has been achieved. Autocorrelation plots graph dependency between draws sampled close together, the presence of which produces downwardly biased estimates of the parameter estimates (Lynch, 2007). It is recommended that several convergence diagnostics be used to determine whether the Markov chains have converged to the posterior distribution.

There are multiple practical decisions to be considered when using MCMC sampling methods in order to ensure convergence to the posterior distribution. Early iterations of Markov chains are often unstable because a large degree of autocorrelation exists between early draws. Researchers normally discard early draws until draws become independent of one another (Kaplan, 2014). This is referred to as the burn-in period. Second, draws taken close to one another will also demonstrate a high degree of autocorrelation. Researchers may combat this problem by thinning the chain by only taking draws from every *xth* iteration. Additionally, it is possible to specify multiple chains from different starting values. Doing so may reduce the number of iterations needed to converge

to the posterior distribution and may help to overcome the effect of poor start values (Kaplan, 2014).

Assessing hypotheses.

Using MCMC methods, Bayesian inference can be synthesized in three steps: establishing the posterior distributions for the parameters of interest for a given model, drawing samples from these posterior distributions via MCMC methods, and summarizing these samples to provide an estimate of the parameter characteristics (Lynch, 2007). Assuming convergence towards the posterior distribution, several methods for testing hypotheses exist (Kaplan, 2014; Lynch, 2007).

A popular method for testing hypotheses is to make inferences from summaries of the posterior distribution. When testing several means against a criterion, two procedures for making inferences are examining the posterior credible interval or specifying a dichotomous parameter for indicating the difference between a level-two mean and the criterion and examining its posterior distribution (Kaplan, 2014; Kruschke, 2013; Nashimoto & Wright, 2008). The second method is discussed in more detail later in this chapter.

The posterior credible interval is a simple method for assessing hypothesis tests involving means, or any parameter for that matter. The posterior credible interval is summarized through quantiles sampled from the posterior distribution (Kaplan, 2014). For instance, the 95% credible interval for a parameter is simply the 2.5 and 97.5 percentiles of the posterior distribution. It is important to note the difference between the Bayesian credible interval and the frequentist

confidence interval. Inferences drawn from the 95% credible interval suggest that 95% of the posterior distribution of the parameter of interest is captured within that interval given the data (Kaplan, 2014). In contrast, the inferences drawn from frequentist confidence interval suggest to the researcher that if it were possible to take an infinite number of independent samples of size *n* from a population, 95% of those samples would contain the parameter of interest (Kaplan, 2014).

In order to determine whether level-two means differ significantly from a criterion, the researcher first derives the desired credible interval for each level-two mean (Gelman et al., 2012; Krushke, 2013). If the credible interval for a given level-two mean excludes the criterion, there is evidence that the two mean is significantly different from the criterion (Kaplan, 2014). Likewise, if the posterior credible interval for a given mean includes the criterion, then there is not sufficient evidence to declare that level-two unit mean as significantly different from the criterion.

Type I error rate.

The above method for assessing hypotheses conducts separate tests for each level-two unit. As a result, the concept of the Bayesian Type I error rate and the issue of multiplicity must be discussed. As with all hypothesis tests, it is possible to arrive at incorrect inferences when conducting a Bayesian hypothesis test. However, while several authors have noted that the Bayesian hypothesis tests tend to be more conservative than their frequentist counterparts, it is generally unknown what the expected Type I error rate for Bayesian tests is *a priori* (Bayarri & Berger, 2004; Gelman et al., 2012; Wang, Leung, Li, & Tan,

2005). As in the frequentist perspective it is desirable that the percentage of Type I errors among a family of tests needs be held under some criterion that is pre-determined by the researcher.

The concept of multiplicity arises solely from the frequentist perspective to statistics. Some Bayesians argue that an adjustment for multiplicity is inherent to Bayesian hypothesis testing (Bayarri & Berger, 2004). As a result, several authors have suggested that in the Bayesian framework there is no need to adjust for multiplicity provided the assumption of conditional independence between hypotheses is met and the prior distribution is correctly specified (Berry & Hochberg, 1999; Gelman et al., 2012; Westfall, Johnson, & Utts, 1997). However, the first assumption is rarely met in practice. Berry and Hochberg (1999) give the humorous example that this assumption might be met when "treatments are of very different types (one a fertilizer and another a human cancer drug, say)" (p. 219). Additionally, testing several means simultaneously likely violates the exchangeability assumption because each mean necessarily refers to a specific category or group (Kaplan, 2014). For example, when discussing several schools, academic achievement at school A cannot be exchanged indiscriminately with the achievement at school B.

Although the idea of multiplicity is a frequentist concept, Westfall et al. (1997) listed several instances in which correction for multiplicity may be necessary from the Bayesian perspective. In particular, the authors state corrections for multiplicity are necessary when the researcher suspects that many of the null hypotheses are true and when the researcher is interested in testing

several hypotheses simultaneously rather than an omnibus test (Berry & Hochberg, 1999; Westfall et al., 1997). As an example, consider a situation in which schools will not be accredited if the average student achievement for a school falls below a criterion identifying the very lowest performing schools. A null hypothesis might be H_0: $\mu_j \leq criterion$. In such a situation, it is expected that the majority of schools will either demonstrate an average performance above the cut point or display an average performance that is statistically indistinguishable from the cut point (i.e., the credible interval for a school contains the cut off value).

Finally, when non-informative priors are assigned the results from Bayesian analyses will approximate the inferences drawn from frequentist analyses because the prior distribution provides relatively little information to the posterior distribution relative to the likelihood (Bayarri & Berger, 2004; Mossman & Berger, 2001). Bayarri and Berger (2004, p. 63) write "The standard normal linear model is the prototypical example: frequentist estimates and confidence intervals coincide exactly with the standard objective Bayesian estimates and credible intervals." Following this logic, when testing hypotheses about multiple means, conducting simultaneous Bayesian *t*-tests will likely lead to conclusions that are in accordance with the increase in Type I errors in their frequentist counterparts due to multiplicity.

Bayesian Approaches to Multiplicity

Acknowledging that it is important to hold the Type I error rate of Bayesian tests under a predetermined α, particularly when the conditional

independence of hypotheses assumption cannot be met and/or uninformative prior distributions are used, four Bayesian models have been proposed. The first method is to model the data as a fully Bayesian MLM (Gelman et al., 2012). The second method is a fully Bayesian hierarchical model with a semi-informative, mixed prior distribution placed upon a difference parameter, δ_q, which signifies the difference between any level-two mean and the grand mean. The third and fourth methods are modifications of the previous two methods that use sample estimates of the within group variances to assign semi informed priors to σ^2.

Bayesian MLM.

Gelman et al. (2012) suggest that a fully Bayesian MLM could be used to control for multiplicity using many of the same arguments presented above in the section on shrinkage. These authors argue that Bayesian MLM accounts for Type I inflation directly by incorporating multiplicity into the model by specifying a hierarchical structure. Additionally, by drawing the level-two means from a larger distribution, the exchangeability assumption is conditionally met (Kaplan, 2014). Bayesian MLMs have also been suggested for use when conducting subgroup analysis in epidemiological studies (Jones, Ohlssen, Neuenschwander, Racine & Branson, 2011).

The Bayesian MLM differs from the frequentist MLM in two ways. The first is that the level-one and two variance components, σ^2 and τ_{00}, are treated as random parameters and each are assigned prior distributions, whereas in the frequentist MLM estimates these two parameters are typically treated as fixed parameters and estimated via maximum likelihood methods. It should be noted

that frequentist MLMs are actually semi-Bayesian in that V_j^* is drawn from the posterior estimate of the variance of β_{0j}^* (Raudenbush & Bryk, 2002). Second, in the Bayesian MLM, the mean of the level-two error term, now denoted as μ_α, is given a prior distribution as well. This is in contrast to the frequentist approach to MLM which fixes μ_α at some value, usually zero.

Previous research has evaluated the use of Bayesian MLMs as a control for multiplicity. Austin et al. (2001) evaluated the performance of Bayesian MLMs when assessing hospital performance, measured as the risk-adjusted mortality rate, as compared to frequentist methods using empirical data. The authors found that the two approaches produced a moderate amount of disagreement in terms of classifying hospital's risk of patient mortality. Gelman et al. (2012) evaluated the performance of Bayesian MLMs against the use of the Bonferroni procedure from the frequentist perspective. They found that the Bayesian MLM method produced different decisions than did the frequentist approach and led to smaller uncertainty about the parameter estimates. Nashimoto and Wright (2008) applied a Bayesian MLM model to a study investigating the effect of cigarette use on lung capacity. This study was purely an empirical application and did not directly address multiplicity.

Few studies have examined the effectiveness Bayesian MLMs as a solution to the multiplicity problem via simulation study (Jones et al., 2011). Yi et al. (2012) compared the Type I error control and power of a Bayesian MLM against several traditional MCPs. The authors found that the Bayesian MLM exhibited a lower Type I error rate and greater power as compared to the

traditional procedures. However, these studies have been limited in scope and generally have not investigated factors that may influence the performance of the Bayesian MLMs. Ohlssen et al. (2006) conducted a simulation study demonstrating the use of logistic Bayesian MLM but conducted their study outside of the context of Type I error inflation. As such, it is unknown how they will perform compared to frequentist MCPs when a large number of hypotheses are tested or in the presence of variance heterogeneity.

Bayesian model for with a δ parameter.

An alternative method to specifying the Bayesian MLM defined above is to define a hierarchical model that treats each mean sequentially based on a simple rank order assumption (Nashimoto & Wright, 2008; Shang, 2011; Shang et al., 2008). Previous research has examined this model in the context of a one-way *ANOVA* (Nashimoto & Wright, 2008) and a two-way mixed *ANOVA* (Li & Shang, 2015; Shang, 2011; Shang et al., 2008). The rank order assumption of means is tenable when there is reason to suspect that a natural ordering of means exists, for instance it is reasonable to suspect that lighter cigarette smokers will exhibit greater lung capacity as compared to heavier smokers (Nashimoto & Wright, 2008). This method was originally developed to test the pairwise difference among means by first ordering the means sequentially and then reparametrizing the J means such that μ_{j+1} is determined by the sum of the preceding mean, μ_j, plus a difference parameter δ_q (Nashimoto & Wright, 2008). More formally, let β_1 denote the smallest mean and then define each remaining $j-1$ means as

$\beta_j = \beta_1 + \sum_{q=1}^{j-1} \delta_q$ for $2 \leq i \leq j$, (Nashimoto & Wright, 2008). The distribution of level-one scores is defined as:

$$[y_{ij} \mid \beta_1, \sigma^2, \delta_1, \delta_2, ..., \delta_q] \overset{iid}{\sim} N(\beta_1 + \sum_{q=1}^{j-1} \delta_q, \sigma^2). \quad (42)$$

Equation 42 necessitates prior distributions for β_1, σ^2, and δ_q. Previous researchers have assigned β_1 a normal prior with a large variance $N(\mu_\alpha, \tau_0^2)$ (Li & Shang, 2015; Nashimoto & Wright, 2008; Shang, 2011; Shang et al., 2008). Nashimoto and Wright (2008) fixed μ_α equal to the sample mean of the smallest group and fixed τ_0^2 to 10^6. In papers by Shang (2008) and Shang et al. (2008) μ_α was fixed at 0 and τ_0^2 was fixed to 100. Nashimoto and Wright (2008) assigned σ^2 an inverse gamma prior with hyperparameters α_0 and β_0. The prior distribution for σ^2 from the papers by Shang (2008) and Shang et al. (2008) are not included because these authors approached this model from a two-way mixed *ANOVA* perspective that includes additional variance components which are not relevant to this paper.

The difference parameter for each rank ordered mean, δ_q, is assigned a prior distribution that is a mixture of an exponential distribution and a discrete distribution with its entire mass at 0 (Li & Shang, 2015; Nashimoto & Wright 2008; Shang, 2011). This is useful when a simple order restriction is present because this prior restricts the difference between two sequential means to be 0 or non-negative.

To model this approach, let I_A be an indicator of the situation in which $\delta_q \neq 0$ and I_B be an indicator of the situation in which $\delta_q = 0$. The prior distribution for δ_q is then assigned as follows:

$$[\delta_q \mid p_q, \eta_q] = p_q I_B + \Delta I_A, \tag{43}$$

and

$$\Delta I_A = (1 - p_q)\frac{1}{\eta_q}\exp\{-\frac{\delta_q}{\eta_q}\}, \tag{44}$$

where η_q is the variance of the normal prior distribution for δ_q and p_q is the probability that δ_q is equal to 0. Prior distributions are then assigned to the hyperparameters of δ_q: p_q and η_q. The noninformative prior *BETA*(A_0, B_0) is assigned to p_q and the noninformative prior $IG(\alpha_0, \beta_0)$ is assigned to η_q. Assigning p_q a *BETA* prior distribution allows the researcher to include information from previous studies pertaining to the probability that δ_q is 0. When no preference is given to the hypothesis that δ_q equals 0, hyperparameters are chosen to ensure p_q equals .5 through the equality:

$$PR\{\delta_i = 0\} = .5 = \frac{A_0}{A_0 + B_0}. \tag{45}$$

This method was adapted by Li and Shang (2016) to test for all pairwise differences when a simple order restriction is not realistic. The rank order assumption may not be tenable when the level-two units are randomly distributed or in exploratory situations where the researcher has little previous knowledge about the order of the level-two units (Li & Shang, 2015). Li and Shang adapted the prior distribution for δ_q to allow the prior to be a mixture of a point mass

distribution entirely centered at 0 and a normal distribution. This allows for the possibility of negative values of δ_q. The new prior distribution for δ_q can now be written as:

$$[\delta_q \mid p_q, \eta_q] = p_q I_B + \Delta I_A, \tag{46}$$

and

$$\Delta I_A = \frac{1}{\sqrt{2\pi\eta_q}} \exp\{-\frac{\delta_q^2}{2\eta_q}\}. \tag{47}$$

The hyperparameters for δ_q, p_q and η_q, remain the same as above.

After estimating the posterior distribution for each δ_q, the remaining pairwise comparisons among non-sequential means can be made by summing the relevant δ_q posterior distributions. For instance, when testing four group means the difference between the second and fourth group means is found by summing the posterior distributions of δ_2 and δ_3, which represents the difference between the second and third means and the third and fourth means. Li and Shang (2016) found that this method maintained a Type I error rate below .05. However, the authors did not compare this method against any other procedures for testing multiple level-two means.

The above approaches are appropriate when making pairwise comparisons between means. This model can be further modified to test hypotheses comparing means to a criterion. This is done by defining δ_q as the difference between a level-two mean and the criterion that, for this study, is the grand centered mean. Each level-two mean is now estimated as:

$$\beta_j = \gamma_{00} + \delta_q, \tag{48}$$

where each term is defined above. The distribution of level-one scores is defined as:

$$\left[y_{ij}\middle|\gamma_{00},\sigma^2,\delta_1,\delta_2,\ldots,\delta_q\right]\sim N(\gamma_{00}+\textstyle\sum_{q=1}^{j}\delta_q,\sigma^2). \quad (49)$$

The prior distribution for σ^2 remains the same as above. The grand mean, γ_{00}, is assigned the same prior distribution as β_1 - $N(\mu_\alpha,\tau_0^2)$. Finally, δ_q is assigned the same prior distributions in Equations 46 and 47. The hyperparameters for δ_q, p_q and η_q, remain the same as above.

For the sake of completeness, a simple Bonferroni correction may be applied to the prior p_q such that:

$$PR\{\delta_q=0\}=.5^{1/(C-1)}=\frac{\mathrm{A}_0}{\mathrm{A}_0+\mathrm{B}_0}. \quad (50)$$

This is a correction for multiple procedures that was suggested by Westfall et al. (1997) and implemented by Shang (2011), Shang et al. (2008), and Li and Wright (2016). Because in the case of a large number of comparisons it will be nearly impossible to observe a situation in which the non-discrete segment of the prior distribution of p_q will be called upon the procedure is not used in this paper. As a result, practitioners would essentially be assigning a discrete prior distribution with its entire mass at zero.

Semi-informative variances.

The third and fourth proposed methods are adjustments to the previous two methods that acknowledge that variance heterogeneity may be present among the level-two units (Nashimoto & Wright, 2008). To combat this heterogeneity a unique inverse gamma prior distribution is assigned to each σ_j^2 by manipulating the hyperparameters α_0 and β_0 so that the mean of the prior distribution for each

group is equal to the sample variance. By directly accounting for variance heterogeneity through incorporating sample variances, estimates of level-two means should be more accurate. A practical application of this method was presented by Nashimoto and Wright (2008), but this adjustment has not been evaluated empirically.

Summary

In the frequentist paradigm, comparing multiple means presents a concern to researchers because the probability of committing a Type I error is inflated. This error inflation may result in various adverse consequences. Under the ongoing example, in which various schools are compared against a common criterion, multiplicity increases the probability that a school is wrongly flagged as performing differently than the national average. As a result, parents may disproportionately choose to send their children to those schools identified as high performing and avoid those schools identified as low performing (Goldstein & Thomas, 1996). In the United States schools that perform exceptionally well may be awarded financial bonuses while poorly performing schools may face sanctions up to and including the loss of accreditation (Raudenbush & Willms, 1995; Schochet & Chiang, 2013).

MCPs have been developed to control for Type I error rate inflation. However, these procedures are known to be very conservative when variance heterogeneity is present and/or a large number of hypotheses are being tested. Both of these factors may be present when testing hypotheses about several means simultaneously. MLMs may provide a solution. MLMs intrinsically control for

type I error inflation by shifting level-two means towards the aggregated mean through shrinkage. In addition, estimates of level-two means become more accurate as the number of level-two units increase.

The Bayesian approach to MLMs provides additional benefits. First, the concept of Type I error inflation does not apply to the Bayesian approach to statistics and, consequently, may ameliorate the problem of multiplicity. However, it is unknown if Bayesian hypothesis tests hold the Type I error rate at or below α. The power of these Bayesian tests as compared to the frequentist methods is unknown as well. Second, Bayesian MLMs allow for the evaluation of the Type I error rate in the situation in which all the level-two means are equal to one another. Third, Bayesian MLMs allow researchers to specify prior distributions that directly model the presence of variance heterogeneity amongst the level-two units.

Present Study

To the best of the author's knowledge, the performance of Bayesian MLMs as an MCP has not been empirically evaluated against traditional MCPs when variance heterogeneity is present and a large number of hypotheses are being tested. Therefore, the purpose of this study was to determine to what extent do Bayesian MLM methods control for Type I error rate when a large number of hypotheses are being tested and variance heterogeneity is present. Additionally, this study examined the power of these procedures compared to two traditional MCPs, the Hochberg and Tukey HSD procedures, in the scenarios in which

variance heterogeneity is present and a large number of hypotheses are being tested.

CHAPTER III. METHODS

To evaluate the above research questions, a Monte Carlo simulation study was conducted using the software platforms *R* version 3.4.2 (R Core Team, 2017) and OpenBugs version 3.2.3 (Lunn, Spiegelhalter, Thomas, & Best, 2009). This chapter begins with an explanation of the data generation process. Following that, the factors that were manipulated and the rationale for the levels of each factor are detailed, along with a description of the constants used in this study. Finally, the methods for calculating the Type I error rate and the power of the proposed procedures are explained.

Data Generation

Data were generated from a three level hierarchical model in *R*. Although the primary interest of this study was to make inferences about level-two means, a three level model was simulated to allow for the scenario in which there is unexplained covariance between level-two units. Under the running example, this may be conceptualized as the effect of an academic district on a subset of schools' average academic achievement. The generating model for the data was as follows:

$$y_{ijk} = \gamma_{000} + \beta_{ijk}x_{ijk} + e_{ijk} + U_{0jk} + V_{00k}, \quad (51)$$

where x_{ijk} is an indicator variable taking on a value of -1 for half the level-two units and 1 for the remaining level-two units, and β_{ijk} is a level-two predictor that acts as the specified effect size. The generating model in Equation 51 allows all the level-two means to take on non-zero values while maintaining the grand mean, γ_{000}, at zero when testing the power of the procedures. Consequently, the

data are generated as grand mean centered. The effect size values and the rationale for the choice of those values are discussed in more detail below. The distributions of the error term for each level were simulated using the RNORM function, which generates random numbers from the normal distribution. The RNORM function requires the user to specify the mean and the standard deviation for each error term. The mean for each error term was set to zero. The values chosen for the standard deviations of the error terms are described in more detail below. Each combination of conditions was simulated 500 times. To ensure replicability the simulation seed was set to 1987 which initialized the random number generator.

Factors

Procedures.

As stated in the previous chapter, six methods for controlling for Type I error inflation are considered: Hochberg's procedure (HOCH), Tukey's HSD (HSD), Bayesian one-way *ANOVA* (B1), Bayesian one-way *ANOVA* with semi-informed variance priors (B1V), Bayesian one-way *ANOVA* with a mean difference parameter (B1D), and Bayesian one-way *ANOVA* with a mean difference parameter and semi-informed variance priors (B1DV). The two traditional MCPs were implemented using pre-existing functions in *R*. These functions correspond to the procedures outlined in Chapter Two and are available in *R*'s base package.

The remaining four methods were modeled in OpenBugs via the *R* package R2OpenBugs (Sturtz, Ligges, & Gelman, 2005). R2OpenBugs translates

the *R* programming language to the syntax used in OpenBugs. Each model is then run in OpenBugs and the results are transferred to the *R* console. This allows for the post processing of the results in the *R* environment.

Level-two sample size.

The number of level-two units was manipulated to act as a proxy for increasing the number of hypotheses that were to be tested. Recall that the number of level-two units is analogous to the number of groups under consideration in the fixed effects approach. A larger number of level-two units correspond to a larger number of testable hypotheses. Manipulating the number of level-two units allows for the investigation of the performance of these six procedures when testing families of hypotheses of differing sizes.

The number of level-two units was chosen to correspond with recommendations for the minimum number of level-two units needed to accurately estimate MLMs. Previous research has suggested that the minimum level-two sample size necessary to provide accurate estimates all parameters in a frequentist MLM ranges between 20 and 50 (Kreft, 1996; McNeish & Stapleton, 2016; Maas & Hox, 2005; Snijders & Bosker, 2012). If the researcher is primarily concerned with accurate estimates of the fixed effects, level-two sample sizes as small as 15 units may produce unbiased results (Baldwin & Fellingham, 2013; McNeish & Stapleton, 2016; Stegmueller, 2013). Bayesian models, which do not carry the frequentist requirement of large sample sizes to obtain asymptotically unbiased estimates, may produce accurate parameter estimates with smaller level-two sample sizes than their frequentist counterparts (Hox, van

de Schoot, & Matthijsse, 2012; McNeish & Stapleton, 2016, Raudenbush & Bryk, 2002; Stegmueller, 2013). Studies have demonstrated that Bayesian MLMs may produce unbiased parameter estimates with level-two sample sizes as small as ten (Austin, 2010; Stegmueller, 2013).

Following these guidelines, level-two sample sizes of 20, and 40 were chosen. An additional level-two sample size of ten was chosen to specifically investigate the performance of Bayesian MLMs when the level-two sample size is small and to provide a condition that corresponds to scenarios in which traditional MCPs have been evaluated in the literature (Donoghue, 1998; Olejnik et al., 1997; Ramsey, 2002). The specific values of the number of level-two units were chosen to allow the remaining factors to be divided equally among the level-two units.

Level-one/within unit variance heterogeneity.

Two conditions were considered for the level-one variance, σ_{ijk}^2. In the first condition, all within group variances were set equal to one. In the second condition, half of the level-one variances were set to .5 and the remaining level-one were set to 1.5. These values were chosen so that the average level-one variance across all level-two units was equal to one for both conditions. Additionally, the values chosen in the heterogeneous condition correspond to moderate variance heterogeneity conditions used in previous research (Kromrey & La Rocca, 1995).

Effect size.

Effect size was defined as the value chosen for the level-two predictor, β_{ijk}. Three levels of effect size were simulated. The first corresponded to the scenario

in which the level-two means did not differ from the grand mean. In this situation, β_{ijk} was set equal to zero. The remaining levels corresponded to the scenario in which the means of the level-two units differed from the criterion of zero. Values of .2 and .5 were chosen for the non-zero values of β_{ijk}. These values were chosen in accordance with Cohen's (1988) recommendations for small and medium effects. Cohen's recommendations are not without criticism. Some have argued that the lack of scale associated with Cohen's effect size recommendations make them inappropriate for some fields. However, because the unadjusted method for testing several means with a criterion corresponds with conducting multiple single, sample *t*-tests Cohen's measure of effect sizes were deemed appropriate.

Level-two *ICC*.

The amount of variation in the dependent variable due to variance at level-two was manipulated via the *ICC2*. Three levels were chosen. An *ICC2* of zero was chosen to simulate the condition in which there is no level-two variance. The remaining two levels were set to .15 and .25, which correspond to values of the *ICC2* seen in the educational research literature (Hedges & Hedberg, 2007a; Hedges & Hedberg, 2007b).

Level-three *ICC*.

The *ICC3* was manipulated to simulate the condition in which unexplained covariance is present among the level-two units due to variation in the third level. Two levels were chosen. The first level corresponds to the scenario in which there is no unexplained covariance among the level-two units. This corresponds

to an *ICC3* of zero. Because research has suggested that the *ICC3* is typically less than the *ICC2* (Siddiqui et. al, 1996), a value of .1 was chosen for the remaining level. This value was consistent with values of the *ICC3* reported by Siddiqui et al. (1996). This level was only simulated in the condition in which the *ICC2* was non-zero. This is because when the *ICC2* is zero there is no covariance among the level-two units and, as a result, the *ICC3* is necessarily zero as well.

Constants

Level-one sample size.

The number of level-one units per level-two unit was fixed in this study. A level-one sample size of 30 was chosen because, by convention, this is the minimum sample size to conduct *t*-tests with adequate power to correctly reject the null hypothesis (Gravetter & Wallnau, 2017).

Level-three sample size.

There is no consensus on the acceptable level-three sample size in the literature. Gelman and Hill (2007) note that, at the bare minimum, two units at the highest level are needed to conduct multilevel analysis with multiple lower level units per higher-level unit. In practice, the number of higher-level units tend to be small due to lack of resources or logistical concerns (Murray, 1998; Donner & Klar, 2000). Fazzari, Kim, and Heo (2014) found, that for certain combinations of *ICC*s and sample sizes of lower level units, level-three sample sizes as small as three may be useable. Applied researchers have analyzed data with as few as ten level-three units (Cunningham, 2010; Grandes, Sanchez, Sanchez-Pinilla, Torcal, Montoya, Lizarraga, & Serra, 2009). As with the level-two sample size, larger

level-three sample sizes are most important when estimating variance components and parameters at the third level, i.e., V_{00k} (Snijders, 2005). Because the focus of this study is on estimating level-two means, the precise estimate of V_{00k} for each replication is not of import. Further, the choice of the level-three sample size is limited by the level-two sample size; for the smallest level-two sample size condition, ten, the largest possible level-three sample size is fixed at five.

Research has found that the level-two sample size per level-three unit has a negligible effect on parameter estimation (Cunningham, 2010; Snijders, 2005). Consequently, the level-three sample size was held constant across all conditions of the level-two sample size. Moreover, since the level-three sample size is limited to five by the smallest level-two condition, a level-three sample size of five was chosen for this study.

Level-two variance.

The level-two error variance, τ^2_{U0}, is dependent on the values of the *ICC2* and *ICC3*. When the *ICC2* was equal to zero, $\tau^2_{U0}$was zero for all levels of *ICC3*, reflecting the lack of variation at level-two. For *ICC2* values of .15 and. 25 and when the *ICC3* was set to zero, $\tau^2_{U0}$is equal to.177, and .333. When the *ICC3* was set to .1, $\tau^2_{U0}$is equal to.159 and .3 for the *ICC2* levels of .15 and .25, respectively. These values are summarized in Table 4.

Table 4. *Values of* τ^2_{U0}

		ICC2		
ICC3		0	0.15	0.25
	0	0	0.17647	0.33333
	0.1	x	0.15885	0.29997

Level-three variance.

As with the level-two error variance, the level-three error variance, τ^2_{V00}, depends on the values of the *ICC2* and *ICC3*. When the *ICC3* was set to zero, $\tau^2_{V00}$was zero for all levels of the *ICC2*. When the *ICC3* was set to .1, $\tau^2_{V00}$was equal to .0175 when the *ICC2* was set to .15, and was equal to .0333 when the *ICC2* was set to .25. These values are summarized in Table 5.

Table 5. *Values of* τ^2_{V00}

		ICC2		
ICC3		0	0.15	0.25
	0	0	0	0
	0.1	x	0.01765	0.03333

Bayesian specifications.

Before conducting Bayesian analysis, the researcher must specify several conditions under which the analysis will be run. Specifically, a decision must be made on which MCMC algorithm to use, the number of chains used in the MCMC walk, the method or methods by which convergence will be monitored, the number of iterations drawn from the MCMC, the burn in period, and the extent to which the MCMC chain(s) will be thinned.

The first decision to be made is which MCMC algorithm to use. For this study, the Gibbs sampler will be used due to its desirable properties when sampling from hierarchical distributions (Lynch, 2007) and because it is the default algorithm used in the R2OpenBugs package in *R* and in OpenBugs (Sturtz et al., 2005). Additionally, the Gibbs sampler has been used in several articles that model or simulate from a Bayesian hierarchical structure (Li & Shang, 2015; Nashimoto & Wright, 2008; Shang, 2011; Shang et al., 2008)

Second, a decision must be made on how many chains to use in the MCMC process. Recall, that using multiple chains can speed the process of convergence. However, each chain requires its own starting values and increases the amount of time necessary to run a simulation. A pilot study demonstrated that two MCMC chains were sufficient to result in acceptable convergence while completing a simulation in a reasonable amount of time.

Third, a decision must be made on how many draws or iterations are to be taken from a parameter's posterior distribution using the MCMC sampling algorithm. The choice of the number of iterations should ensure that the MCMC algorithm converged to a tractable solution and, as a result, a method for convergence must be selected as well. As stated in Chapter Two, there are several methods for assessing convergence; two of which are through inspecting trace plots and the use of the $\hat{R}$ test statistic. However, because it is not feasible to visually inspect every replication, the $\hat{R}$ statistic will be used to assess convergence in this study. For *m* number of chains, each with a length of *d*, it is possible to calculate the mean of a given parameter for each chain ($\bar{\theta}_j$), the aggregate mean of the parameter over all chains ($\bar{\theta}$), the within chain (V_{wc}) and the between chain variance (V_{bc}). V_{wc} is calculated as:

$$[(1/(m(\mathrm{d}-1))\sum_{i=1}^{m}\sum_{j=1}^{d}(\theta_{ij}-\theta_{i})^2], \tag{52}$$

and V_{bc} is calculated as:

$$[\mathrm{d}/(\mathrm{m}-1)\sum_{i=1}^{m}(\bar{\theta}_i-\bar{\theta})^2]. \tag{53}$$

The total variance of a parameter, V_{tot}, can then be calculated as:

$$(d-1)/dV_{wc}+(1/d)V_{bc}. \tag{54}$$

$\hat{R}$ is calculated as V_{tot}/V_{wc}. Values less than 1.1 provide evidence that the MCMC procedure has reached convergence while values of $\hat{R}$ greater than 1.5 provide the researcher with considerable doubt about the validity of the estimates drawn from the posterior distribution (Gelman et al., 2012). For the B1 and B1V procedures, pilot testing demonstrated that 3000 iterations were sufficient to hold

$\hat{R}$ below 1.1 for all parameters in a model for the vast majority of replications. Unfortunately, pilot testing also demonstrated evidence that the B1D and B1DV models rarely produced $\hat{R}$ values less than 1.1. As a result, a more lenient convergence criterion of 1.5 was applied to the B1D and B1DV models. Replications that did not keep $\hat{R}$ below 1.1, for the B1 and B1V procedures, or 1.5, for the B1D and B1DV procedures, were discarded and the model was rerun with a larger number of replications until the criterion was met.

The default number of initial draws that were discarded (the burn-in period) used in R2OpenBugs was chosen for this study. The length of the burn-in period is determined in R2OpenBugs by dividing the number of iterations by two. The R2OpenBug's default thinning process was also used. R2OpenBugs determines this by dividing the difference between the number of iterations and the length of the burn in period by 1000 and then multiplying that ratio by the number of chains.

Prior distributions.

Bayesian procedures necessitate that a prior distribution be chosen for each variable in the model. All parameters were given uninformative prior distributions. Recalling that the four procedures assume a two level hierarchical structure, all four models share the random variable γ_{00}. The aggregate level-two mean is assigned its own normal, prior distribution with hyperparameters μ_α and τ_{00}, which are the mean and variance, respectively, of the normal prior distribution. By drawing each level-two mean from a common prior distribution,

the likelihood of each level-two mean has an influence on the posterior distribution of every other level-two mean.

Because MLMs assume a hierarchical structure, the hyperparameters for γ_{00} are also assigned either fixed values or prior distributions. The hyperparameter μ_α is fixed to zero for the B1D and B1DV models. This is a reasonable decision, because it corresponds to the scenario in which academic achievement has been grand mean centered. The hyperparameter μ_α is assigned the prior distribution, *N(0,100)*, for the B1 and B1V models. This prior again is reasonable in light of the running scenario. The prior distribution for μ_α was assigned a large variance to make it extremely non-informative. τ_{00} is assigned the prior distribution *U*[.0001 ,100] for all four models. The rationale for this specific prior distribution for τ_{00} will be discussed in the next paragraph.

The four of the Bayesian models also share that random variable σ^2. As with τ_{00}, the parameter σ^2 as assigned the prior distribution *U*[.0001, 100]. The uniform distribution was chosen based on work by Gelman (2006) who recommended the use of a uniform prior distribution for both τ_{00} and σ^2 when one is first beginning an iterative process of fitting a MLM or the researcher is not particularly interested in selecting a conjugate prior. Because we have elected to select prior distributions that are as non-informative as possible, the second recommendation seems to apply. The two hyperparameters of the uniform distribution represent the lower limit, A, and upper limit, B, respectively. To justify the selected values of A and B, recall that variance parameters are bounded by zero and infinity. Focusing on B, a value need be selected that is large enough

to encapsulate all plausible values of the sample variance. Beyond this threshold, which is unknown, the choice of the hyperparameter B becomes arbitrary. Because the only requirement to ensure the prior distributions for σ^2 and τ_{00} are uninformative is that B is sufficiently large, a value of 100 was chosen based on an applied example given by Gelman (2006). The value of A was chosen so that neither variance component would be estimated as having a negative value. However, OpenBugs requires the specification of the precision of the normal distribution rather than the variance. The precision is simply the inverse of the variance. As a result, a value of zero could not be used as a lower bound for the variance parameters, and it was necessary to add a small constant to the lower bound of the prior distribution.

The B1V and B1DV models allow σ_j^2 to vary across level-two units. This parameter is still assigned a uniform distribution. However, now it is distributed as $U[.0001, \mathrm{B}_j]$, where B_j is chosen so that the mean of the prior distribution is equal to the sample variance of each level-two mean. The mean of a uniform distribution can be found by $\frac{\mathrm{A}+\mathrm{B}}{2}$. Substituting the level-two sample variance, s_j^2, for the mean of the uniform distribution and rearranging terms, we can solve for each B_j by:

$$\mathrm{B}_j = 2s_j^2 - A. \tag{55}$$

The B1D and B1DV models contain three additional parameters that necessitate prior distributions. The difference parameter, δ_q, is assigned a prior

distribution that is a mixture of a point mass distribution with its entire mass at zero and a normal distribution (Li & Shang, 2015). The distribution is written as:

$$[\delta_q \mid p_q, \eta_q] = p_q I_B + \Delta I_A, \tag{56}$$

and

$$\Delta I_A = \frac{1}{\sqrt{2\pi\eta_q}} \exp\{-\frac{\delta_q^2}{2\eta_q}\}. \tag{57}$$

Following the suggestion by Li and Shang (2016) the hyperparameters for δ_q, p_q and η_q, are assigned the prior distributions *BETA*(1, 2) and *IG*(2.1, .0005).

Outcomes

Type I error.

The Type I error rate is calculated only in the condition in which the effect size, β_{ijk}, was simulated to be to zero. When evaluating the Type I error of these procedures, the null hypothesis tested whether each level-two mean was different from the criterion, which was also zero. A Type I error is identified as occurring when a *p*-value less than or equal to .05 is observed for the Tukey or Hochberg procedures or, for the Bayesian procedures, when the 95% credible intervals for any level-two mean excludes zero. The Type I error rate was defined as the proportion of replications for a given set of conditions in which at least one Type I error was identified.

The B1V and The B1VD models allow an alternative method for assessing whether a Type I error occurred. Recall that the parameter δ_q, a parameter representing the difference between a level-two mean and the aggregate mean, can take on a value of either zero or any other non-zero number. One may declare a level-two mean to be significantly different from the criterion by inspecting the

posterior distribution of δ_q and determining whether more than half the values in the posterior distribution take on a value other than zero. If this proportion is greater than .5 evidence exists that the two level-two mean under consideration is different from the criterion (Li & Shang, 2015; Nashimoto & Wright, 2008; Shang et al., 2008;). A type I error would occur when more than half the values of the posterior distribution of δ_q take on a non-zero value under any of the conditions in which β_{ijk} equals zero. The average Type I error rate can be defined as the number of replications in which the posterior probability of δ_q was greater than .5 divided by the total number of replications for a given set of conditions, when all level-two means are set equal.

Table 6. *Type I Error Criteria*

	HSD & HOCH	Bayesian Models	B1D & B1DV
Type I Error	p <= .05	95% credible intervals excludes 0	> 50% of the posterior distribution for δ_q is non-zero

Power.

The all-pair definition of power to reject all false null hypotheses in a family of tests was used in this study. The power of these procedures is evaluated in the conditions in which β_{ijk} was set to either .2 or .5. The power of each procedure was measured by identifying those hypotheses in which a difference between two means was correctly detected. A correct decision occurred when a *p*-value less than or equal to .05 is observed for the Tukey or Hochberg procedures or, for the Bayesian procedures, when the credible intervals for any two means exclude one another. The power per replication was defined as the proportion of correct decisions among all hypotheses. The average power was

defined as the mean power for each procedure across all replications for a given set of conditions.

As with the Type I error rate, the power of the B1V and B1VD models can be defined with respect to the posterior distribution of δ_q. Specifically, the power for a family of hypotheses can be defined as the proportion of hypotheses in which more than half the values of the posterior mean of δ_q take on a value greater than 0 for any level-two mean. As is the case above, the power per replication was defined as the proportion of correct decisions among all hypotheses for a given set of conditions with average power defined as the mean power for each procedure across all replications for a given set of conditions.

Table 7. *Power Criteria*

	HSD & HOCH	Bayesian Models	B1D & B1DV
Power	p <= .05	95% credible interval excludes zero	> 50% of the posterior distribution for δ_q is non-zero

CHAPTER IV. RESULTS

A Monte Carlo simulation was conducted to evaluate the performance of the six procedures across the data generation settings described in Chapter Three. The data generation settings are summarized in Table 8. When the *ICC3* was specified to be non-zero, these methods were only evaluated under the condition in which τ^2_{U0} was also non-zero. This decision was made due to the following reasons.

First, the presence of level-three variance would be equivalent to adding a constant to the level-two mean. Analyzing various values of the level-two mean is a condition that is already investigated in this study. To demonstrate this, consider the situation in which one wished to estimate the mean of the first level-two unit, nested within the first level-three unit, $\bar{y}_{.11}$. Further, suppose γ_{000} is specified to be zero, β_{i11} is specified to be .2, e_{ijk} is expected to sum to zero, the level-three random effect, V_{00k}, is taken from Table 5 and distributed as *N(0, .01765)*, and no level-two random effect is present. Using the above parameters and the generating Equation 51, $\bar{y}_{.11}$ becomes equal to $\beta_{i11} + V_{001}$. Second, it is not possible to generate the data to have a pre-specified *ICC3* if the level-two variance is zero. By inspecting Equation 33, it can be seen that in such a scenario the *ICC3* will always be equal to 1. Finally, a data structure with variance at level-three and no variance at level-two lacks real world plausibility. If such a structure was found, it is likely that the data would be analyzed using a two level MLM with the level-three units treated as the second level grouping factor. As a result of this decision, 90 generation conditions were simulated.

Table 8. *Data Generation Conditions*

$\sigma^2_{ijk} = 1$				$\sigma^2_{i1:N/2k} = 0.5$ *and* $\sigma^2_{i(\frac{N}{2})+1:Nk} = 1.5$			
β_{ijk}	*ICC3*	*ICC2*	*N*	β_{ijk}	*ICC3*	*ICC2*	*N*
0	0	0	10	0	0	0	10
			20				20
			40				40
		0.15	10			0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40
	0.1	0.15	10		0.1	0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40
0.2	0	0	10	0.2	0	0	10
			20				20
			40				40
		0.15	10			0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40
	0.1	0.15	10		0.1	0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40
0.5	0	0	10	0.5	0	0	10
			20				20
			40				40
		0.15	10			0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40
	0.1	0.15	10		0.1	0.15	10
			20				20
			40				40
		0.25	10			0.25	10
			20				20
			40				40

Chapter Four is structured as follows. First, a summary of the data simulation process is presented in the data generation section. In the nonconvergence section, the process of ensuring model convergence for the Bayesian methods is discussed. The Type I error control and power of the six methods across all simulation conditions are presented in the primary analysis section along with a comparison of the six methods under consideration.

Data Generation Summary

The data were generated using *R* from the three level multilevel model specified in Equation 51. The model was manipulated to generate the data under the specified simulation conditions. For each combination of simulation settings, 500 replications were drawn. The simulation settings are the expected values of the parameters for the three level MLM for a given combination of simulation settings. To assess whether the data were generated as specified, four indices were considered: the mean value of the parameter across replications, the parameter bias, the standard deviation (*SD*) of the parameter estimate, and root mean squared error (*RMSE*) of the parameter estimate.

The mean value of the parameter across replications was found by summing each parameter estimate in a given combination of simulation conditions and dividing by the number of replications. The bias was found by:

$$bias = \sum_{r=1}^{1000} \frac{\hat{\theta}}{1000} - \theta, \tag{58}$$

where θ is the simulation parameter, $\hat{\theta}$ is the generated parameter estimate, and *r* indexes the simulation replication. The *SD* of the parameter estimate was found by:

$$SD = \sqrt{\frac{(\hat{\theta} - \sum_{r=1}^{1000} \frac{\hat{\theta}}{1000})^2}{1000}}. \tag{59}$$

The *RMSE* of the parameter, which can be conceptualized as an index of the precision of the estimated parameter as compared to its generating parameter, was found through the equation:

$$RMSE = \sqrt{bias^2 + SD^2}, \quad (60)$$

where the bias and *SD* of the parameter are defined in Equations 58 and 59. To facilitate discussion of the parameters, the results were divided between the estimates of the mean parameters (γ_{000} and β_{ijk}), variance parameters (σ^2_{ijk}, τ^2_{u0k}, and τ^2_{V00}), and the *ICC* (*ICC2* and *ICC3*).

Mean estimates.

The mean estimates of the data generation model consisted of the grand intercept, γ_{000}, and the level-two effect, β_{ijk}. To obtain a clear view of the accuracy of the data generation process, these values are presented under the conditions in which the level-two and three variance components, and consequently the *ICC2* and *ICC3*, were specified to be zero. Table 9 contains the data generation results for the mean parameters when all level-one variances were set equal to one and Table 10 contains the data generation results for the mean parameters when half the level-two within group variances were specified to be .5 and the remaining half were specified to be 1.5. A full table of the mean data generation results for all simulation conditions can be found in Appendix C.

Table 9. *Mean Parameter Generation Results for all* $\sigma^2_{ijk} = 1$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.002	0.001	0.001	0	0.002	0	0	0	0.001
	Bias	0.002	0.001	0.001	0	0.002	0	0	0	0.001
	SD	0.058	0.039	0.029	0.059	0.041	0.029	0.058	0.04	0.029
	RMSE	0.058	0.039	0.029	0.059	0.041	0.029	0.058	0.04	0.029
$\widehat{\beta_{ijk}}$	Mean	0.002	0.001	0.001	0.196	0.199	0.199	0.499	0.5	0.499
	Bias	0.002	0.001	0.001	-0.004	-0.001	-0.001	-0.001	0	-0.001
	SD	0.058	0.039	0.029	0.057	0.042	0.029	0.058	0.04	0.029
	RMSE	0.058	0.039	0.029	0.059	0.041	0.029	0.058	0.04	0.029

Table 10. *Mean Parameter Generation Results for* $\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0	-0.001	-0.001	0	0	0	-0.001	0.001	-0.001
	Bias	0	-0.001	-0.001	0	0	0	-0.001	0.001	-0.001
	SD	0.057	0.04	0.028	0.06	0.041	0.029	0.057	0.042	0.029
	RMSE	0.057	0.04	0.029	0.06	0.041	0.029	0.057	0.042	0.029
$\widehat{\beta_{ijk}}$	Mean	0	-0.001	-0.001	0.2	0.199	0.2	0.498	0.502	0.5
	Bias	0	-0.001	-0.001	0	-0.001	0	-0.002	0.002	0
	SD	0.057	0.04	0.028	0.061	0.04	0.029	0.057	0.041	0.028
	RMSE	0.057	0.04	0.029	0.06	0.041	0.029	0.057	0.042	0.029

As can be seen in Tables 9 and 10, the estimated parameters were close in value to the expected parameters. The precision of the estimated parameters varied as a function of the level-two sample size, *N*. As the level-two sample size increased, the simulated mean parameters became more similar to the expected values specified in the simulation model. Because the level-one sample size was set to 30 for each level to unit, this indicates that the simulated mean values became more accurate as the total sample size increased (given that the total sample size can be found by the product of *N* and 30).

Variance estimates.

The variance estimates of the data generation model consisted of the level-one variance, σ^2_{ijk}, level-two variance, τ^2_{u0k}, and the level-three variance, τ^2_{V00}. To obtain an unadulterated view of the accuracy of the data generation process, these values are presented in the simulation condition in which the level-two mean, β_{ijk}, is specified to be zero. When β_{ijk} is non-zero, the variance components are influenced by the added effect and no longer correspond to their generating parameters. A full table of the variance data generation results for all simulation conditions can be found in Appendix C. Tables 11-14 present the data generation results for the variance estimates.

Table 11. *Variance Generation Results for all* $\sigma^2_{ijk} = 1$ and $ICC3 = 0$

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	0.996	0.994	0.996	0.996	1.004	1.001	1	1	0.999
	Bias	-0.004	-0.006	-0.004	-0.004	0.004	0.001	0	0	-0.001
	SD	0.085	0.058	0.04	0.081	0.06	0.042	0.082	0.059	0.041
	RMSE	0.085	0.058	0.04	0.081	0.06	0.042	0.082	0.059	0.041
$\widehat{\tau^2_{u0k}}$	Mean	0.004	0.004	0.003	0.141	0.163	0.17	0.268	0.307	0.328
	Bias	0.004	0.004	0.003	-0.036	-0.014	-0.007	-0.065	-0.026	-0.005
	SD	0.008	0.007	0.004	0.099	0.071	0.047	0.171	0.122	0.079
	RMSE	0.009	0.008	0.005	0.105	0.072	0.047	0.183	0.125	0.08
$\widehat{\tau^2_{V00}}$	Mean	0.003	0.002	0.001	0.036	0.015	0.008	0.064	0.027	0.013
	Bias	0.003	0.002	0.001	0.036	0.015	0.008	0.064	0.027	0.013
	SD	0.008	0.004	0.002	0.059	0.028	0.015	0.106	0.049	0.023
	RMSE	0.009	0.004	0.002	0.069	0.032	0.017	0.124	0.056	0.027

Table 12. *Variance Generation Results for all* $\sigma^2_{ijk} = 1$ and *ICC3*= .1

	τ^2_{u0k}	0			0.159			0.3		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	x	x	x	0.997	1.003	1	1.001	0.999	1.002
	Bias	x	x	x	-0.003	0.003	0	0.001	-0.001	0.002
	SD	x	x	x	0.086	0.056	0.041	0.085	0.057	0.041
	RMSE	x	x	x	0.086	0.057	0.041	0.085	0.057	0.041
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.132	0.149	0.158	0.252	0.279	0.295
	Bias	x	x	x	-0.027	-0.01	-0.001	-0.048	-0.021	-0.005
	SD	x	x	x	0.125	0.144	0.155	0.15	0.145	0.145
	RMSE	x	x	x	0.128	0.145	0.155	0.157	0.146	0.145
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.047	0.028	0.021	0.085	0.051	0.036
	Bias	x	x	x	0.029	0.011	0.003	0.052	0.018	0.003
	SD	x	x	x	0.094	0.106	0.11	0.078	0.046	0.031
	RMSE	x	x	x	0.098	0.106	0.11	0.094	0.049	0.031

Table 13. *Variance Generation Results for* $\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$ and $ICC3 = 0$

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	0.501	0.501	0.501	0.499	0.5	0.5	0.502	0.501	0.499
	Bias	0.001	0.001	0.001	-0.001	0	0	0.002	0.001	-0.001
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.001	0.001	0.001	0.001	0	0	0.002	0.001	0.001
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	1.497	1.505	1.502	1.502	1.504	1.502	1.496	1.495	1.502
	Bias	-0.003	0.005	0.002	0.002	0.004	0.002	-0.004	-0.005	0.002
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.003	0.005	0.002	0.002	0.004	0.002	0.004	0.005	0.002
$\widehat{\tau^2_{u0k}}$	Mean	0.004	0.004	0.003	0.142	0.163	0.17	0.27	0.317	0.316
	Bias	0.004	0.004	0.003	-0.034	-0.013	-0.007	-0.063	-0.016	-0.018
	SD	0.01	0.007	0.005	0.094	0.067	0.048	0.172	0.12	0.081
	RMSE	0.011	0.008	0.006	0.101	0.068	0.049	0.183	0.121	0.083
$\widehat{\tau^2_{V00}}$	Mean	0.004	0.002	0.001	0.037	0.015	0.008	0.064	0.027	0.014
	Bias	0.004	0.002	0.001	0.037	0.015	0.008	0.064	0.027	0.014
	SD	0.009	0.004	0.002	0.064	0.028	0.015	0.105	0.048	0.025
	RMSE	0.01	0.005	0.003	0.074	0.032	0.017	0.123	0.055	0.028

Table 14. *Variance Generation Results for* $\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$ and $ICC3 = .1$

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	x	x	x	0.498	0.501	0.498	0.502	0.5	0.501
	Bias	x	x	x	-0.002	0.001	-0.002	0.002	0	0.001
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0.002	0.001	0.002	0.002	0	0.001
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	x	x	x	1.49	1.504	1.498	1.505	1.497	1.499
	Bias	x	x	x	-0.01	0.004	-0.002	0.005	-0.003	-0.001
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0.01	0.004	0.002	0.005	0.003	0.001
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.138	0.152	0.155	0.26	0.286	0.298
	Bias	x	x	x	-0.021	-0.007	-0.004	-0.04	-0.014	-0.002
	SD	x	x	x	0.13	0.147	0.152	0.154	0.149	0.149
	RMSE	x	x	x	0.132	0.148	0.152	0.159	0.15	0.149
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.044	0.028	0.022	0.082	0.052	0.04
	Bias	x	x	x	0.027	0.01	0.004	0.049	0.019	0.007
	SD	x	x	x	0.099	0.109	0.104	0.081	0.048	0.032
	RMSE	x	x	x	0.103	0.11	0.104	0.095	0.051	0.033

The level-one variance estimates closely approximated their generation parameters for all combinations of sample size, *ICC2*, and *ICC3*. These close approximations held when half the level-one variances were specified to take on values of .5 and the remaining half were specified to take on values of 1.5.

The precision of τ^2_{u0k}, on the other hand, varied as a function of the sample size, the *ICC2*, and the *ICC3*. Unequal level-one variances has a negligible impact on the estimates of τ^2_{u0k} as compared to the condition in which all level-one variance were specified to be equal to one. With two exceptions, the estimates of the level-two variance became more precise as *N* increased. These exceptions occurred when the *ICC2* was specified to be .15 and the *ICC3* was specified to be 0.1 for both the equal and unequal level-one variance conditions. In this scenario, the precision of the simulated τ^2_{u0k}

actually decreased as *N* increased. However, this was due to the standard deviation of the parameter estimate increasing with sample size; the bias of τ^2_{u0k} decreased as the sample size increased.

Across all conditions, the precision of τ^2_{u0k} decreased as the expected value of τ^2_{u0k} increased. The effect of the expected value of τ^2_{u0k} on the generated values of τ^2_{u0k} interacted with the sample size. Increasing the level-two sample size mitigated the effects of simulating larger values of τ^2_{u0k} to some extent. This is consistent with the literature discussed in Chapter Three, where evidence was provided that the higher-level variance components become more precise as the corresponding higher-level sample size increases.

The generated values of τ^2_{u0k} tended to become less precise as the level-three variance, and consequently the *ICC3*, increased. This is likely due to covariance in the level-three unit adding additional covariance at level-two. Increasing the sample size did not have a consistent effect on the precision of τ^2_{u0k} as τ^2_{V00} was increased.

The precision of the level-three variance, τ^2_{V00}, was primarily influenced by the total sample size. Unequal level-one variances has a negligible impact on the estimates of τ^2_{V00} as compared to the condition in which all level-one variance were specified to be equal to one. Compared to the simulated values of τ^2_{u0k}, the simulated values of τ^2_{V00} were much poorer estimates of their generating parameter. This is likely due to the level-three sample size being held constant at five for all simulation conditions. The literature cited in Chapter Three noted that level-three variance estimates tend to be imprecise when the level-three sample size is small. Increasing the expected value of τ^2_{u0k} or τ^2_{V00}, via the *ICC3*, did not have a consistent linear effect on the precision of τ^2_{V00}.

With two exceptions, the estimates of the level-three variance became more precise as the total sample size increased. These exceptions occurred when the *ICC2* was specified to be .15 and the *ICC3* was specified to be 0.1 for both the equal and unequal level-one variance conditions. In this scenario, the precision of the simulated τ^2_{V00} actually decreased as the sample size increased. However, this was due to the standard deviation of the parameter estimate differing in a non-monotonic fashion with sample size; the bias of τ^2_{V00} decreased as the sample size increased.

***ICC* estimates.**

The *ICC* estimates of the data generation model consisted of the parameters for the *ICC2* and *ICC3*. The estimated values of the *ICC2* and *ICC3* were found by applying Equations 32 and 33 to the relevant variance estimates. To obtain an unbiased view of the accuracy of the data generation process, these values are presented in the simulation condition in which β_{ijk} is specified to be zero. When β_{ijk} is non-zero, the variance components are influenced by the added effect mean effect and no longer correspond to their generating parameters. Consequently, the *ICC* estimates will no longer correspond to their generating parameters. A full table of the *ICC* data generation results for all simulation conditions can be found in Appendix C. Tables 15-18 present the data generation results for the variance estimates.

Table 15. *ICC Generation Results for all* $\sigma_{ijk}^2 = 1$ and $ICC_3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.007	0.006	0.004	0.145	0.148	0.15	0.238	0.245	0.252
	Bias	0.007	0.006	0.004	-0.005	-0.002	0	-0.012	-0.005	0.002
	SD	0	0	0	0.002	0.002	0.001	0.003	0.002	0.001
	RMSE	0.007	0.006	0.004	0.005	0.002	0.001	0.012	0.006	0.003
$\widehat{ICC3}$	Mean	0.336	0.334	0.33	0.206	0.087	0.047	0.194	0.079	0.039
	Bias	0.336	0.334	0.33	0.206	0.087	0.047	0.194	0.079	0.039
	SD	0.453	0.443	0.44	0.292	0.145	0.078	0.273	0.133	0.065
	RMSE	0.564	0.555	0.551	0.357	0.169	0.091	0.334	0.155	0.076

Table 16. *ICC Generation Results for all* $\sigma_{ijk}^2 = 1$ and $ICC_3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.148	0.148	0.15	0.24	0.243	0.246
	Bias	x	x	x	-0.002	-0.002	0	-0.01	-0.007	-0.004
	SD	x	x	x	0.003	0.003	0.003	0.004	0.003	0.003
	RMSE	x	x	x	0.004	0.004	0.003	0.01	0.008	0.005
$\widehat{ICC3}$	Mean	x	x	x	0.254	0.148	0.109	0.239	0.139	0.103
	Bias	x	x	x	0.154	0.048	0.009	0.139	0.039	0.003
	SD	x	x	x	0.663	0.791	0.846	0.311	0.19	0.127
	RMSE	x	x	x	0.681	0.792	0.846	0.341	0.194	0.127

Table 17. *ICC Generation Results for* $\sigma_{i1:N/2k}^2 = .5$ and $\sigma_{i\left(\frac{N}{2}\right)+1:Nk}^2 = 1.5$ and $ICC_3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.008	0.006	0.004	0.147	0.149	0.149	0.24	0.251	0.245
	Bias	0.008	0.006	0.004	-0.003	-0.001	-0.001	-0.01	0.001	-0.005
	SD	0	0	0	0.002	0.002	0.001	0.003	0.002	0.001
	RMSE	0.008	0.006	0.004	0.004	0.002	0.001	0.011	0.002	0.005
$\widehat{ICC3}$	Mean	0.349	0.319	0.313	0.203	0.085	0.042	0.19	0.077	0.041
	Bias	0.349	0.319	0.313	0.203	0.085	0.042	0.19	0.077	0.041
	SD	0.459	0.441	0.434	0.29	0.144	0.077	0.266	0.128	0.07
	RMSE	0.576	0.544	0.536	0.354	0.167	0.088	0.328	0.149	0.081

Table 18. *ICC Generation Results for* $\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$ and $ICC_3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.15	0.149	0.149	0.243	0.247	0.249
	Bias	x	x	x	0	-0.001	-0.001	-0.007	-0.003	-0.001
	SD	x	x	x	0.003	0.003	0.003	0.004	0.003	0.003
	RMSE	x	x	x	0.003	0.003	0.003	0.008	0.004	0.003
$\widehat{ICC3}$	Mean	x	x	x	0.232	0.142	0.116	0.235	0.14	0.109
	Bias	x	x	x	0.132	0.042	0.016	0.135	0.04	0.009
	SD	x	x	x	0.68	0.797	0.839	0.301	0.181	0.126
	RMSE	x	x	x	0.693	0.798	0.839	0.33	0.186	0.127

The estimates of the simulated values of the *ICC2* were more robust to variations in the generating factors than were the variance estimates. The generated values of the *ICC2* were influenced by the total sample size and the expected *ICC2*. Unequal level-one variances had a negligible impact on the estimates of the *ICC2* as compared to the condition in which all level-one variance were specified to be equal to one. Additionally, increasing the expected value of the *ICC3* had a negligible impact on the precision of the simulated *ICC2*.

Across all simulation conditions, the estimates of the *ICC2* became more precise as the total sample size increased. This was expected due to the relationship between the variance components and the *ICC2*. When the *ICC3* was specified to be .1, the precision of the *ICC2* became worse as the expected value of the *ICC2* increased. This relationship did not hold when the *ICC3* was specified to be zero.

The estimates of the *ICC3* were less precise that the estimates of the *ICC2*. This was expected due to the greater influence of the level-three variance in calculating the *ICC2* along with the effect of the small level-three sample size on the precision of the level-three variance. The estimates of the *ICC3* were influenced by the sample size, the

expected value of the *ICC2* and the expected value of the *ICC3*. Unequal level-one variances had a negligible impact on the estimates of the *ICC3* as compared to the condition in which all level-one variance were specified to be equal to one.

In general, the precision of the simulated *ICC3* increased as the total sample size increased. A notable exception to this pattern was when the expected value of the *ICC3* was .1 and the expected value of the *ICC2* was .15. In these generating scenarios, the standard error of the simulated *ICC3* increased with the total sample size.

As the expected value of the *ICC2* increased, the precision of the simulated *ICC3* decreased across all conditions. The precision of the *ICC3* was worse when the expected value of the *ICC3* was .1 as compared to when it was specified to be zero. This relationship held across all conditions.

Data generation summary.

In general, the simulated parameters provided an adequate approximation of their generating parameters. The estimated mean parameters, estimated level-one variance, and estimated *ICC2* produced a low *RMSE* across simulation conditions.

The exceptions to this trend were the higher-level variance parameters and the *ICC3*. The simulated value of τ^2_{u0k} was generally a precise estimate of its generating parameter when no level-three variance was specified. When the level-three variance was specified to be non-zero, the precision of the simulated τ^2_{u0k} became noticeably worse. The generated values of τ^2_{V00} and the *ICC3* lacked precision particularly when the total sample size was small and when the expected *ICC2* was small. However, the poor precision of τ^2_{V00} and the *ICC3* was expected due to the small level-three sample size. Additionally, the multilevel methods for controlling Type I error inflation are all two

level models, so the precision of the level-three estimates did not have a meaningful effect on the results.

Nonconvergence

As discussed in Chapter Three, it is essential to ensure that the posterior distributions of the parameters converged to an admissible solution. In order to evaluate whether the Bayesian models achieved convergence, the $\hat{R}$ estimates were investigated for each replication. Values of $\hat{R}$ less than 1.1 indicate that the researcher may have confidence that that the posterior distribution of the model converged correctly (Gelman el al., 2012). Values of $\hat{R}$ greater than 1.5 provide the researcher with considerable doubt about the validity of the estimates drawn from the posterior distribution (Gelman el al., 2012). Values of $\hat{R}$ less than 1.1 were obtained for the B1 and B1V models for the majority of replications under the Bayesian settings specified in Chapter Three. Replications that produced a $\hat{R}$ value greater than 1.1 for any parameter from the B1 and B1V models were discarded. Additional replications, with a larger number of draws taken from the posterior distribution, were generated to ensure that all simulation settings were composed of 500 replications that met the convergence criterion. All parameters from all replications of the B1 and B1V model met the convergence criterion of 1.1 when 5,000 draws were taken from the relevant posterior distribution.

The parameters of the B1D and B1DV models, on the other hand, were much more likely to produce a $\hat{R}$ greater than 1.1 than the B1 and B1V models. Increasing the number of draws from the posterior distribution of these models substantially would likely have ameliorated the situation. The majority of parameters from the B1D and B1DV models produced corresponding $\hat{R}$ values that were less than 1.5. Replications

that produced a $\hat{R}$ value greater than 1.5 for any parameter from the B1D and B1DV models were discarded. Additional replications, with a larger number of draws taken from the posterior distribution of each parameter, were generated to ensure that all simulation settings were composed of 500 replications that met the convergence criterion. All parameters from all replications of the B1 and B1V model met the convergence criterion of 1.5 when 12,000 draws were taken from the relevant posterior distribution.

Primary Analysis

The primary analysis of this paper consisted of evaluating the Type I error control and power of six methods for controlling for multiplicity: Hochberg's procedure (HOCH), Tukey's HSD (HSD), Bayesian one-way *ANOVA* (B1), Bayesian one-way *ANOVA* with semi-informed variance priors (B1V), Bayesian one-way *ANOVA* with a mean difference parameter (B1D), and Bayesian one-way *ANOVA* with a mean difference parameter and semi-informed variance priors (B1DV). The Type I error rate and power of these procedures were evaluated under each of the 90 simulation conditions. Relevant tables and figures are provided when appropriate.

Type I error rate.

The Type I error rate of the six procedures was evaluated in the condition in which β_{ijk} was specified to be zero. A Type I error was identified as occurring when a *p*-value less than or equal to .05 was observed for the HSD or HOCH procedures or, for the Bayesian procedures, when the 95% credible intervals for any level-two mean excluded zero – this will be referred to as the *traditional Type I error rate*. For the B1D and B1DV procedures, a Type I error was additionally defined as occurring when more than half the values of the posterior distribution of δ_q take on a non-zero value – this will be referred to

as the *alternative Type I error rate*. Both definitions of a Type I error were used when evaluating the B1D and B1DV models. Additionally, the unadjusted Type I error rate was found by conducting single sample *t*-tests on each of the level-two means and comparing the p-values obtained from those tests to α. The Type I error rate was defined as the proportion of replications for a given set of conditions in which at least one Type I error was identified – this corresponds to the definition of the familywise Type I error rate. The section on the Type I error rate of the six procedures under the various simulation conditions is organized as follows. First, the main effects of varying the level-two sample size are discussed. The remaining factors are then discussed it terms of their effect on the Type I error rate across the three levels of *N*. The unadjusted Type I error rate is found in Table 19 and the Type I error rate of the six procedures are found in Tables 20.

Table 19. *Unadjusted Type I Error Rates*

		all $\sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$		
ICC3	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 20	*N* = 40
0	0	.387	.665	.866	.419	.67	.859
	0.15	.992	1	1	1	1	1
	0.25	.999	1	1	1	1	1
0.1	0.15	.997	1	1	.996	1	1
	0.25	.999	1	1	1	1	1

Table 20. *Adjusted Type I Error Rates*

			$all\ \sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5\ and\ \sigma^2_{i(\frac{N}{2})+1:Nk} = 1.5$		
Method	*ICC3*	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 10	*N* = 40
HOCH	0	0	.05	.063	.047	.053	.048	.043
HSD			.008	.003	.002	.015	.01	.005
B1			.084	.062	.034	.12	.104	.076
B1V			.074	.05	.018	.08	.072	.032
B1D			0(0)	0(.016)	.002(.014)	0(.01)	.006(.06)	.006(.088)
B1DV			0(.002)	0(.028)	0(.028)	0(.004)	0 (.012)	0(.014)
HOCH	0	0.15	.91	.989	.999	.969	996	1
HSD			.675	.83	.937	.6	.739	.893
B1			.976	.998	1	.986	1	1
B1V			.952	.998	1	1	1	1
B1D			.498(.756)	.738(.912)	.962(1)	.514(.77)	.764(.954)	.97(1)
B1DV			.378(.684)	.696(.934)	.91(.994)	.582(.814)	.818(.964)	.988(1)
HOCH	0	0.25	.983	.999	1	.995	1	1
HSD			.836	.948	.993	.781	.944	.992
B1			.994	1	1	1	1	1
B1V			.984	1	1	1	1	1
B1D			.786(.912)	.962(1)	1(1)	.8(948)	.978(.996)	1(1)
B1DV			.73(.918)	.928(1)	1(1)	.84(.95)	.97(.998)	1(1)
HOCH	0.1	0.15	.933	.985	999	.961	.999	.999
HSD			.67	.799	.901	.614	.748	.879
B1			.976	1	1	.97	1	1
B1V			.968	.994	1	.984	1	1
B1D			.514(.764)	.772(.954)	.94(.982)	.46(.752)	.756(.948)	.972(.998)
B1DV			.39(.692)	.676(.952)	.876(.996)	.538(.792)	.836(.97)	.974(1)
HOCH	0.1	0.25	.999	1	1	.997	1	1
HSD			.828	.951	.987	.799	.928	.986
B1			.996	1	1	1	1	1
B1V			.996	1	1	.998	1	1
B1D			.798(.926)	.97(1)	.994(.998)	.814(.96)	.978(.996)	.998(1)
B1DV			.732(.908)	.934(.986)	1(1)	.842(.956)	.986(1)	1(1)

Note: Values in parentheses indicate the alternative Type I error rate.

Level-two sample size.

In order to obtain the sharpest view of the effect of varying the level-two sample size, the Type I error rate of the procedures is first discussed under the condition in which all level-one variances were equal to one and the *ICC2* and *ICC3* were specified to be zero. The unadjusted Type I error increased from 0.387 when *N* was equal to 10, to 0.665 when *N* was equal to 20 and to 0.866 when *N* was equal to 40. These Type I error rates approximately correspond to the expected familywise Type I error rate defined in Equation 3.

Increasing the level-two sample size generally decreased the Type I error rate of the six procedures, resulting in more conservative procedures. Among the traditional MCPs, the HSD procedure was able to maintain strong control of the Type I error rate for the three levels of level-two sample size. The Type I error rate of the HOCH procedure exceeded α when *N* was equal to 20 but otherwise maintained control of the Type I error rate at or below α.

This result warranted further investigation. By applying Equation 3 to Hochberg's procedure, the expected Type I error rate for 20 hypothesis tests is .0488. To determine if the observed Type I error rate of .063 was a reasonable simulation result, a Wald confidence interval was constructed using Equation 61:

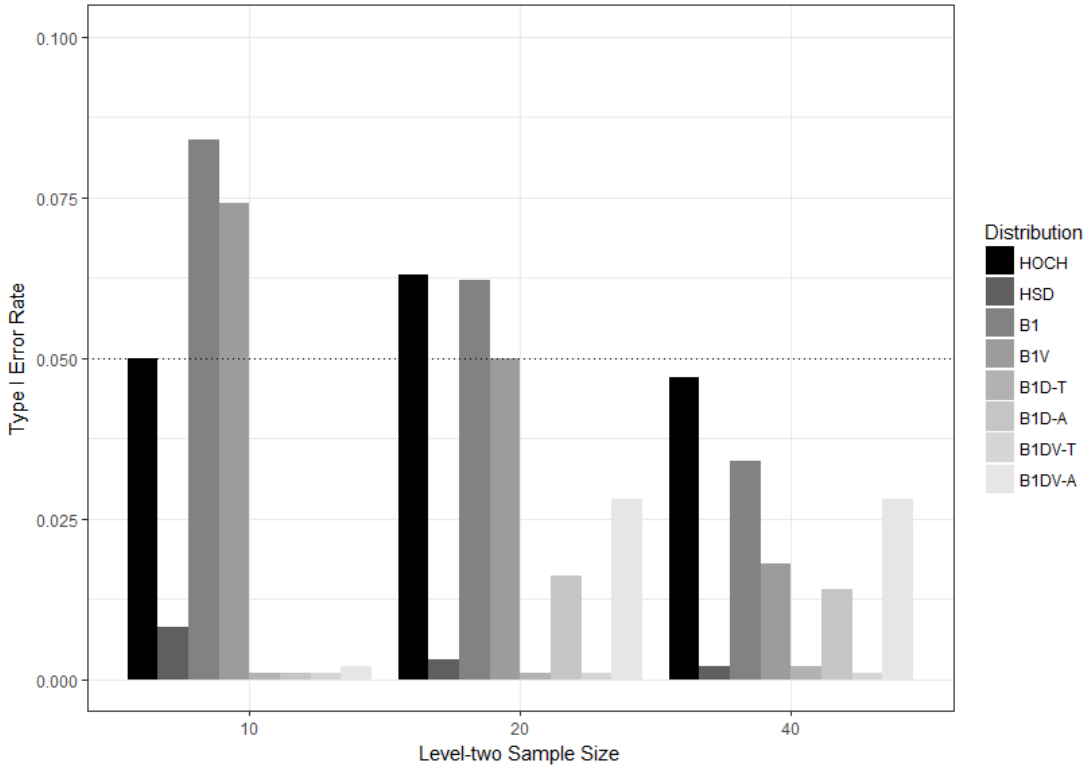
$$CI = \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{r}}, \tag{61}$$

where $\hat{\pi}$ is the observed Type I error rate and *r* is the number of replications. The resulting 95% confidence interval (.0417, .0843) provided evidence that the observed Type I error rate was a realistic value drawn from a sampling distribution with an expected Type I error rate of .0488. Further, changing the simulation seed ameliorated

this issue, leading to the expected pattern of the Type I error rate of the HOCH decreasing as *N* increased.

The Type I error rate of the B1 procedure exceeded α when *N* was equal to 10 and 20 but was held below α when *N* was equal to 40. The Type I error rate of the B1V procedure exceeded α when *N* was equal to 10 and was maintained at or below α when *N* was equal to 20 and 40. Both the B1D and B1DV procedures maintained control of the Type I error below α under both definitions of the Type I error rate.

Figure 2. Type I Error Rate by N (Homogenous Level-One Variances; ICC2 & ICC3 = 0)



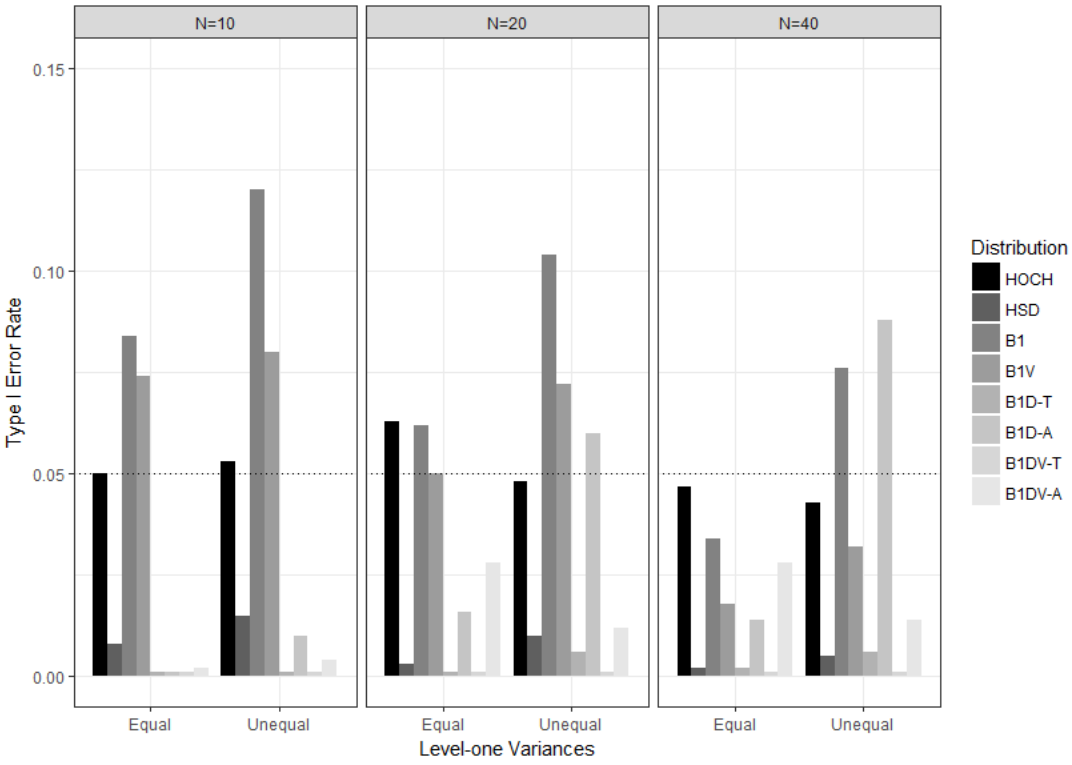
Level-one variance heterogeneity.

The effect of variance heterogeneity among the level-one units was explored in the conditions in which the *ICC2* and *ICC3* were specified to be zero. The presence of

variance heterogeneity among the level-one units slightly increased the unadjusted Type I error rates across all levels of the level-two sample size.

Among the traditional MCPs, the presence of level-one variance heterogeneity increased the Type I error rate of the HSD procedure and had an inconclusive effect on the Type I error rate of the HOCH procedure. In the heterogeneous level-one variance condition, the HSD procedure was able to maintain the Type I error rate below α for all conditions of N while the HOCH procedure was able to maintain the Type I error rate below α when N was equal to 20 and 40 but failed to do so when N was equal to 10. The Bayesian procedures were more adversely affected by heterogeneous level-one variances. The Type I error rate of the B1 and B1V procedures increased when the level-one variance were heterogeneous. The B1 was not able to maintain the Type I error rate at α for any level of N when the level-one variances were unequal while the B1V procedure was only able to maintain the Type one error rate below α when N was equal to 40 under the heterogeneous level-one variance condition. Because the Type I error rates of the B1D and B1DV procedures were so close to zero, for both definitions of the Type I error rate, it was difficult to tease out the effect of heterogeneous level-one variances on the Type I error rate of these procedures. The B1D and B1DV procedures maintained the Type I error rate well below α for both the homogenous and heterogeneous variance conditions.

Figure 3. Type I Error Rate by Level-One Variance Condition (ICC2 & ICC3 = 0)

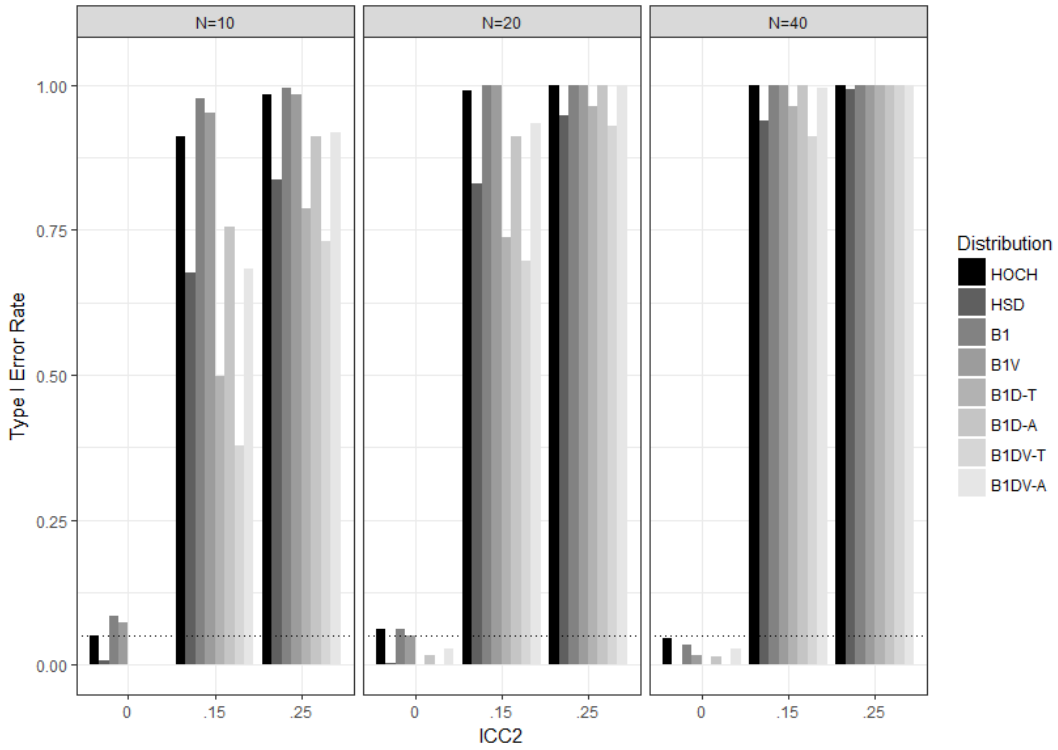


ICC2.

The effect of the interaction of varying non-zero *ICC2* and *N* on the Type I error control of the procedures was analyzed under the condition in which the *ICC3* was specified to be zero. At least one Type I error was recorded for the unadjusted *p*-values across all replications when the *ICC2* was non-zero. The presence of level-two variability, through specifying the *ICC2* to be non-zero, resulted in a Type I error in the majority of replications for all procedures. This is expected however, because, like adding level-three variability to the model as discussed above, the addition of level-two variability is equivalent to adding a mean effect at level-two.

All procedures failed to maintain the Type I error rate at α for every condition in which the *ICC2* was non-zero. As *N* increased, the probability of committing a Type I error increased for all procedures across all non-zero *ICC2* conditions. The probability that a procedure committed a Type I error was greater in the heterogeneous level-one variance condition than in the heterogeneous level-one variance condition for both non-zero values of the *ICC2*.

Figure 4. Type I Error Rate by ICC2 (Homogenous Level-One Variances and ICC3 = 0)



ICC3.

To contextualize the effect of the varying the *ICC3* on the Type I error rate of the six procedures, the results need be discussed across different levels of the *ICC2*. However, non-zero values of the *ICC2* resulted in Type I error rates that greatly exceeded *α*, regardless of whether the *ICC3* was zero or not. As a result, it is difficult to parse out the effect of varying the *ICC3*. In any case, all of the procedures failed to maintain the Type I error rate below *α* when the *ICC3* was non-zero. As *N* increased the probability of committing a Type I error increased. The probability of committing a Type I error was greater in the heterogeneous variance condition than in the homogenous variance condition.

Figure 5. Type I Error Rate by ICC3 (Homogenous Level-One Variances; ICC2 = .15)

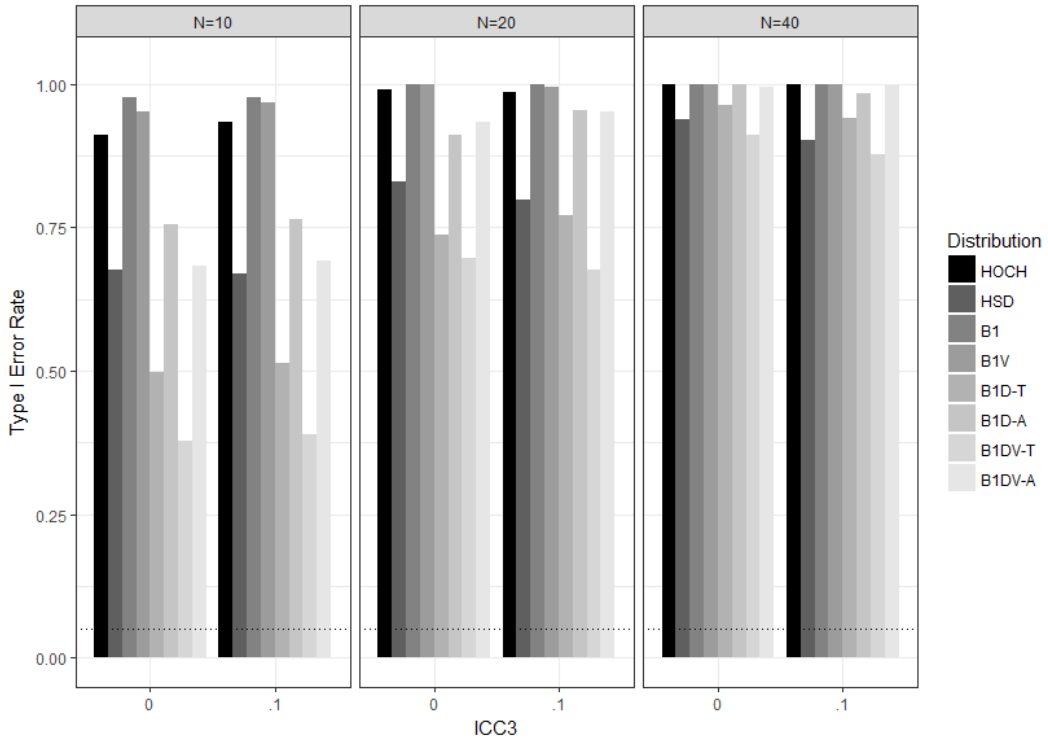
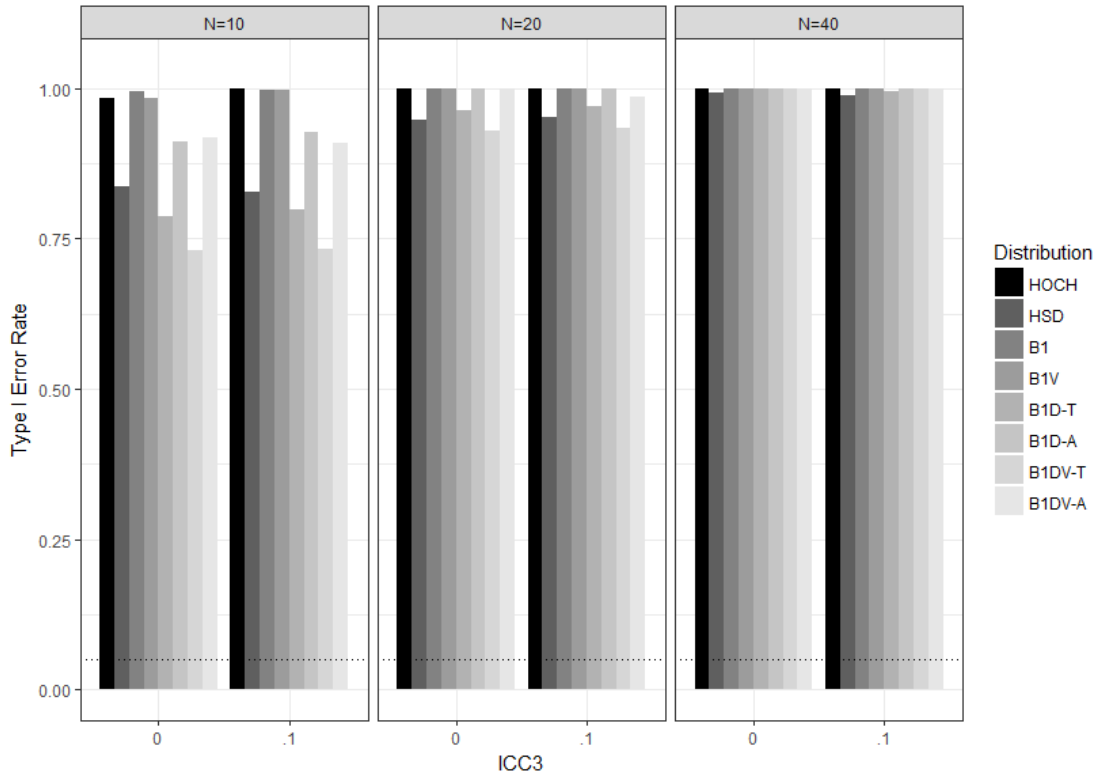


Figure 6. Type I Error Rate by ICC3 (Homogenous Level-One Variances; ICC2 = .25)



Comparison of the procedures.

None of the six procedures was able to maintain strong control of the Type I error rate at α. Recall, a procedure is said to maintain strong control of the Type I error rate if it produces a Type I error rate at or below α for all simulation conditions. The procedures that were able to maintain strong control of the Type I error rate in the "ideal" simulation condition (homogenous level-one variances and *ICC2* and *ICC3* specified to be zero) were the HSD procedure, B1D and B1DV methods. In particular, the B1 and B1V procedures displayed difficulty maintaining the Type I error rate below α for the smaller level-two sample size conditions and when level-one variance heterogeneity was present. When *N* was large, all procedures were able to maintain control of the Type I error rate below α given that there was no variance at level-two or three and that all level-one variances were equal to one another. Non-zero values of the *ICC2* and *ICC3*, on the

other hand, had a massive and negative effect on the Type I error control of the six procedures. The presences of any non-zero level-two and three variances rendered these procedures useless as a control for maintaining the Type I error at α. Generally, for a given simulation condition, the B1 procedure was the most likely procedure to commit a Type I error followed by the B1V, Hoch, HSD, B1D, and B1DV procedures.

Power.

The power of the six procedures was evaluated in the simulation conditions in which β_{ijk} was specified to be non-zero. The power of a procedure is defined as the proportion of sample means that were determined to be significantly different than the aggregate mean, which was grand mean centered at zero, averaged over the number of replications for a given simulation condition. A sample mean is said to be significantly different from zero if its adjusted *p*-value was less than or equal to .05 for the HSD or HOCH procedures or, for the Bayesian procedures, when the 95% credible intervals for any level-two mean excludes zero – this will be referred to as the *traditional power*. For the B1D and B1DV procedures, a Type I error was alternatively defined as occurring when more than half the values of the posterior distribution of δ_q take on a value of zero – this will be referred to as the *alternative power*. Both definitions of a Type I error were used when evaluating the B1D and B1DV models. Additionally, the unadjusted power was found by conducting single sample *t*-tests on each of the level-two means. The section on the power of the six procedures under the various simulation conditions is organized as follows. First, the effect of varying the level-two sample size is discussed. This discussion takes place under the context of different levels of the effect size. The remaining factors are then discussed it terms of their effect across the three levels of the

level-two sample size. The unadjusted power rate are found in Tables 21 and 22. The power of the six procedures are found in Tables 23 and 24.

Table 21. *Unadjusted Power when β_{ijk} = .2*

		all $\sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$		
ICC3	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 20	*N* = 40
0	0	.184	.186	.183	.233	.233	.232
	0.15	.458	.463	.47	.487	.483	.49
	0.25	.566	.563	.563	.585	.582	.581
0.1	0.15	.466	.465	.459	.495	.493	.489
	0.25	.567	.565	.563	.587	.586	.583

Table 22. Unadjusted Power when β_{ijk} = .5

		all $\sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$		
ICC3	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 20	*N* = 40
0	0	.748	.754	.752	.771	.773	.77
	0.15	.64	.634	.641	.654	.655	.657
	0.25	.658	.662	.659	.671	.672	.672
0.1	0.15	.641	.643	.648	.654	.656	.653
	0.25	.653	.655	.66	.673	.681	.675

Table 23. Power of the Six Procedures when β_{ijk} = .2

			all $\sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$		
Method	*ICC3*	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 20	*N* = 40
HOCH	0	0	.039	.023	.012	.059	.036	.023
HSD			.008	.005	.002	.017	.009	.004
B1			.092	.078	.064	.096	.067	.064
B1V			.076	.057	.047	.142	.111	.115
B1D			<.001(.006)	<.001(.011)	.001(.009)	.001(.009)	. 002 (.015)	.002 (.013)
B1DV			.001(.007)	<.001(.007)	.001(.007)	.001(.012)	.001(.01)	.001(.013)
HOCH	0	0.15	.296	.255	.216	.332	.295	.257
HSD			.121	.096	.078	.108	.089	.071
B1			.434	.443	.45	.435	.44	.439
B1V			.403	.414	.421	.435	.442	.445
B1D			.09(.184)	.098(.193)	.111(.2)	.091(.182)	.108(.2)	.111(.201)
B1DV			.069(.161)	.083(.184)	.088(.184)	.108(.197)	.116(.214)	.116(.217)
HOCH	0	0.25	.428	.379	.341	.458	.414	.37
HSD			.184	.157	.135	.16	.14	.121
B1			.561	.565	.562	.55	.56	.56
B1V			.538	.539	.535	.555	.555	.555
B1D			.195(.307)	.203(.308)	.209(.312)	.186(.291)	.202(.306)	.209(.312)
B1DV			.158(.283)	.166 (.286)	.175(.298)	.19(.301)	.199(.315)	.206(.318)
HOCH	0.1	0.15	.301	.259	.213	.344	.294	.256
HSD			.131	.101	.077	.112	.087	.074
B1			.453	.451	.438	.44	.44	.449
B1V			.43	.424	.408	.447	.445	.445
B1D			.1(.188)	.107(.198)	.103(.194)	.091(.184)	.101(.188)	.11(.204)
B1DV			.077(.165)	.077(.177)	.083(.181)	.106(.206)	.103(.201)	.117(.212)
HOCH	0.1	0.25	.431	.381	.334	.459	.419	.369
HSD			.178	.154	.133	.167	.147	.123
B1			.557	.56	.556	.56	.564	.553
B1V			.534	.537	.532	.563	.562	.554
B1D			.177(.278)	.201(.305)	.202(.3)	.193(.299)	.198(.313)	.203(.308)
B1DV			.155(.269)	.166(.285)	.169(.286)	.196(.302)	.2(.319)	.199(.31)

Note: Values in parentheses indicate the alternative power definition.

Table 24. Power of the Six Procedures when β_{ijk} = .5

			all $\sigma^2_{ijk} = 1$			$\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i(\frac{N}{2})+1:Nk} = 1.5$		
Method	*ICC3*	*ICC2*	*N* = 10	*N* = 20	*N* = 40	*N* = 10	*N* = 20	*N* = 40
HOCH	0	0	.511	.372	.265	.617	.524	.423
HSD			.162	.102	.062	.143	.106	.068
B1			.74	.741	.737	.753	.762	.758
B1V			.688	.683	.68	.716	.714	.713
B1D			.036(.176)	.04(.195)	.039(.195)	.039(.17)	.041(.177)	.043(.184)
B1DV			.036(.182)	.038(.198)	.037(.201)	.088(.292)	.092(.309)	.09(.305)
HOCH	0	0.15	.517	.46	.417	.543	.493	.45
HSD			.217	.185	.164	.205	.177	.152
B1			.631	.635	.642	.634	.644	.641
B1V			.61	.61	.618	.635	.633	.639
B1D			.24(.377)	.251(.387)	.257(.393)	.252(.39)	.25(.388)	.257(.388)
B1DV			.204(.359)	.212(.354)	.212(.36)	.254(.392)	.247(.39)	.251(.397)
HOCH	0	0.25	.553	.511	.466	.575	.531	.492
HSD			.242	.218	.191	.223	.2	.182
B1			.656	.663	.663	.66	.659	.662
B1V			.637	.645	.644	.656	.655	.658
B1D			.317(.428)	.314(.43)	.33(.444)	.304(.421)	.313(.428)	.322(.44)
B1DV			.269(.408)	.274(.412)	.282(.421)	.301(.432)	.304(.433)	.313(.44)
HOCH	0.1	0.15	.522	.462	.422	.54	.496	.447
HSD			.22	.186	.165	.2	.175	.152
B1			.637	.637	.639	.635	.637	.63
B1V			.617	.618	.618	.625	.63	.626
B1D			.25(.384)	.25(.379)	.264(.392)	.251(.388)	.247(.381)	.253(.385)
B1DV			.208(.356)	.211(.365)	.211(.362)	.251(.391)	.25(.392)	.244(.388)
HOCH	0.1	0.25	.547	.506	.465	.575	.542	.499
HSD			.238	.216	.193	.225	.206	.184
B1			.65	.652	.662	.663	.662	.665
B1V			.63	.632	.641	.661	.658	.662
B1D			.287(.398)	.322(.43)	.315(.43)	.322(.434)	.319(.436)	.326(.434)
B1DV			.26(.395)	.269(.409)	.278(.418)	.317(.436)	.313(.442)	.314(.446)

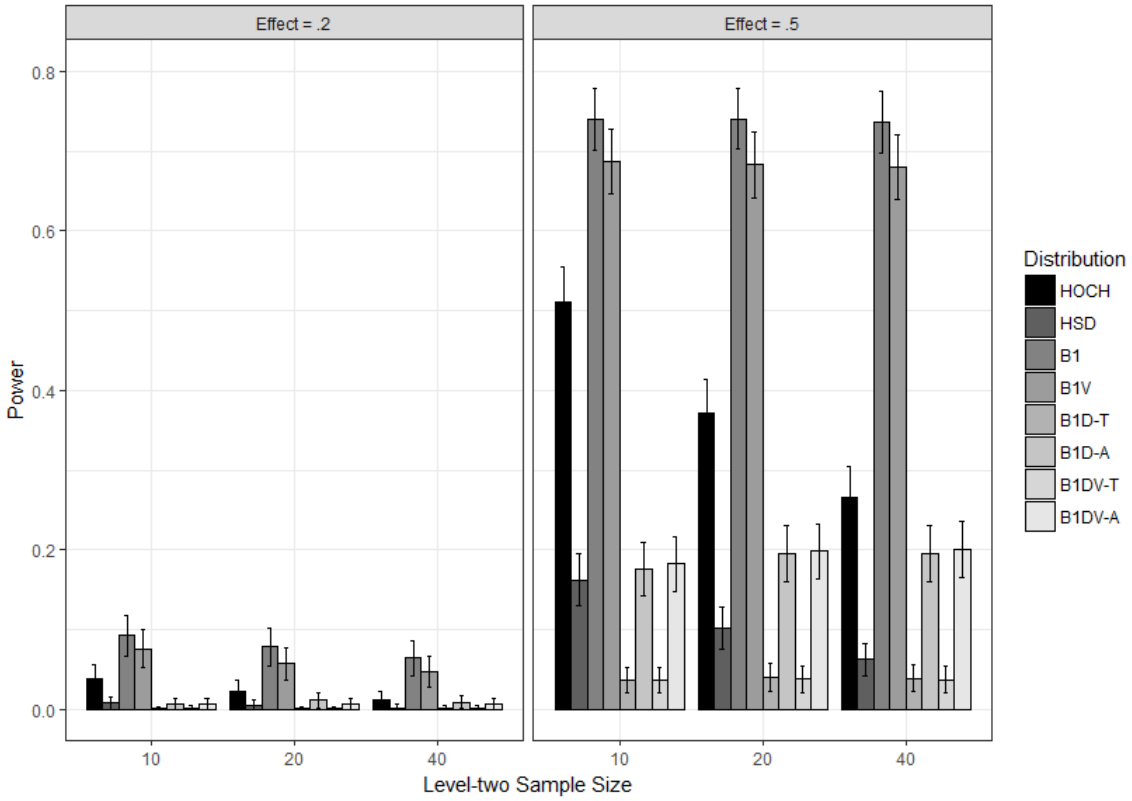
Note: Values in parentheses indicate the alternative power.

Level-two sample size.

The influence of the level-two sample size on the power of the six procedures was evaluated under the conditions in which the level-one variances were homogenous and the *ICC2* and *ICC3* were specified to be zero. The level-two sample size had no effect on the unadjusted power when β_{ijk} was .2 and slightly increased the unadjusted power when β_{ijk} was .5.

The effect of increasing *N* on the six procedures was mixed as well. The power of both the HOCH and HSD procedures decreased as *N* increased. When β_{ijk} was .2, the power of the B1 and B1V procedures decreased as *N* increased. There is evidence that this pattern held for the B1 procedure when β_{ijk} was .5; however, the decrease was less noticeable. The power of the B1V procedure remained relatively unchanged for different values of *N* when β_{ijk} was .5. The traditional power of the B1D and B1DV procedures were close to zero in these two conditions and as a result it was difficult to discern any effect of varying the level-two sample size. The same was true of the alternative power of these two procedures when β_{ijk} was .2. When β_{ijk} was .5, the alternative power of the B1D procedure tended to increase with the level-two sample size. This pattern did not hold for the B1DV procedure.

Figure 7. Power by Effect Size & N (Homogenous Level-One Variances; ICC2=0; ICC3=0)



Effect size.

The influence of the effect size on the power of the six procedures was evaluated under the conditions in which the level-one variances were homogenous and the *ICC2* and *ICC3* were specified to be zero. Unsurprisingly, increasing the effect size resulted in an increase in the unadjusted power and the power of the six procedures. The largest increase in power occurred for the B1 procedure while the HSD, B1D, and B1DV procedures demonstrated the smallest increase in power. The increase in power due to effect size was relatively constant across *N* for all procedures.

ICC2.

Increasing the *ICC2*, resulted in increased power for each of the procedures. In fact, the largest power across all simulation conditions was observed in this scenario for the B1 procedure. As stated above, the effect of setting the *ICC2* to a non-zero value was expected due to the presence of variance at level-two being equivalent to adding a non-zero value to the level-two mean. The effect of varying the *ICC2* was constant over all values of *N* and effect sizes.

Figure 8. Power by ICC2 (Homogenous Level-One Variances; Effect = .2; ICC3 = 0)

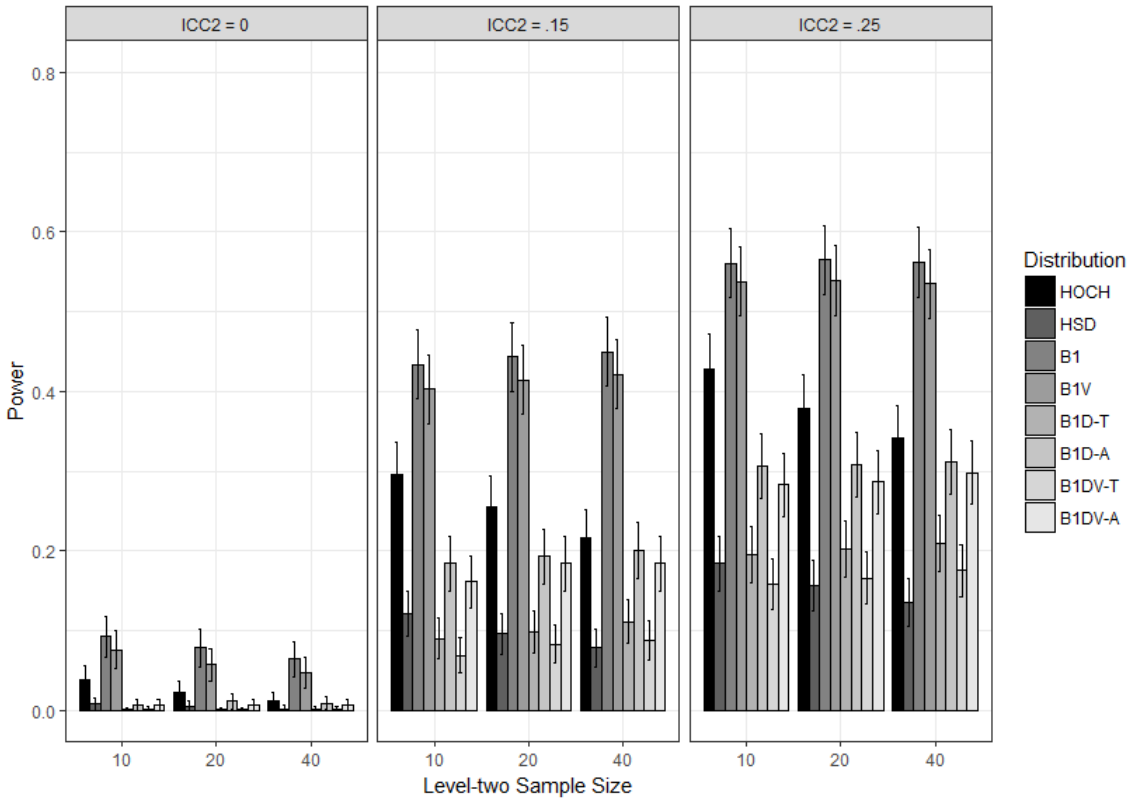
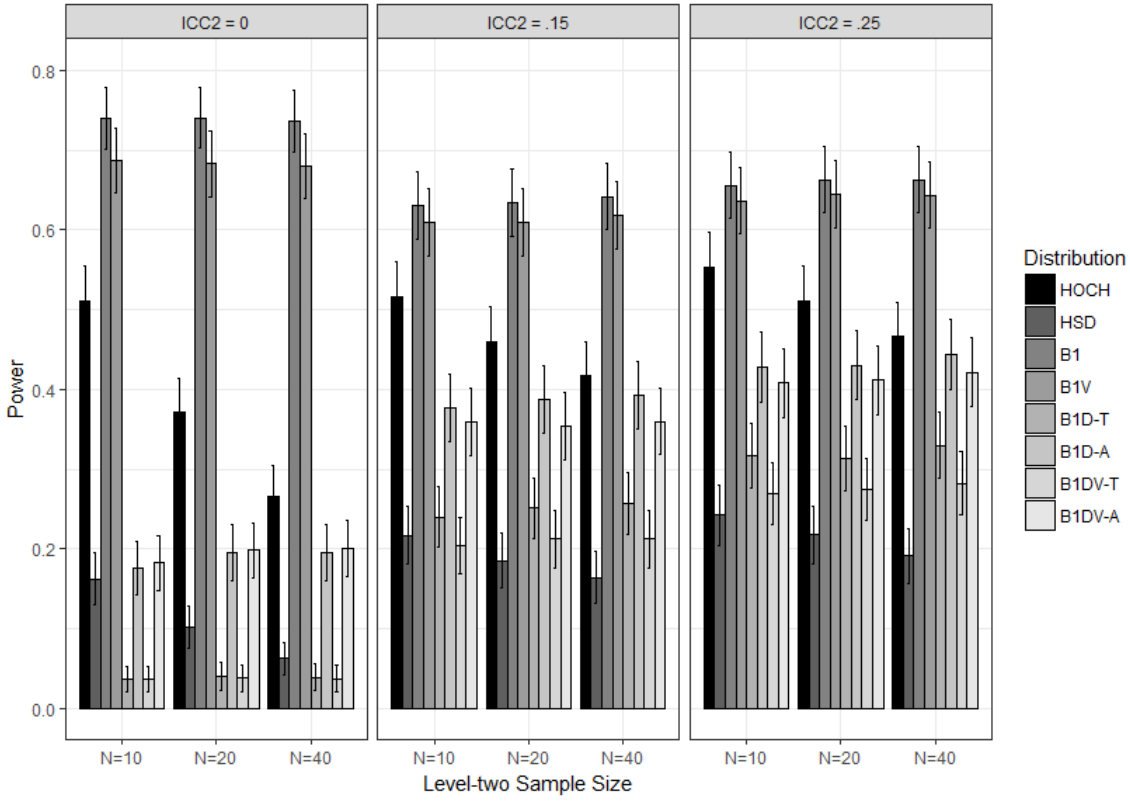


Figure 9. Power by ICC2 (Homogenous Level-One Variances; Effect = .5; ICC3 = 0)



ICC3.

The effect of varying the *ICC3* was dependent on the value of the *ICC2*. When the *ICC2* was specified to be .15, increasing the *ICC3* from zero to .1 resulted in an increase in power for all procedures and across all levels of the effect size an *N*. However, the increase in power was more modest when β_{ijk} was equal to .5 than when β_{ijk} was equal to .2. When the *ICC2* was specified to be .25, increasing the *ICC3* from zero to .1 did not noticeably increase the power of the procedures. Again, this pattern held for all levels of the effect size and *N*.

Figure 10. Power by ICC3 (Homogenous Level-One Variances; Effect = .2; ICC2 = .15)

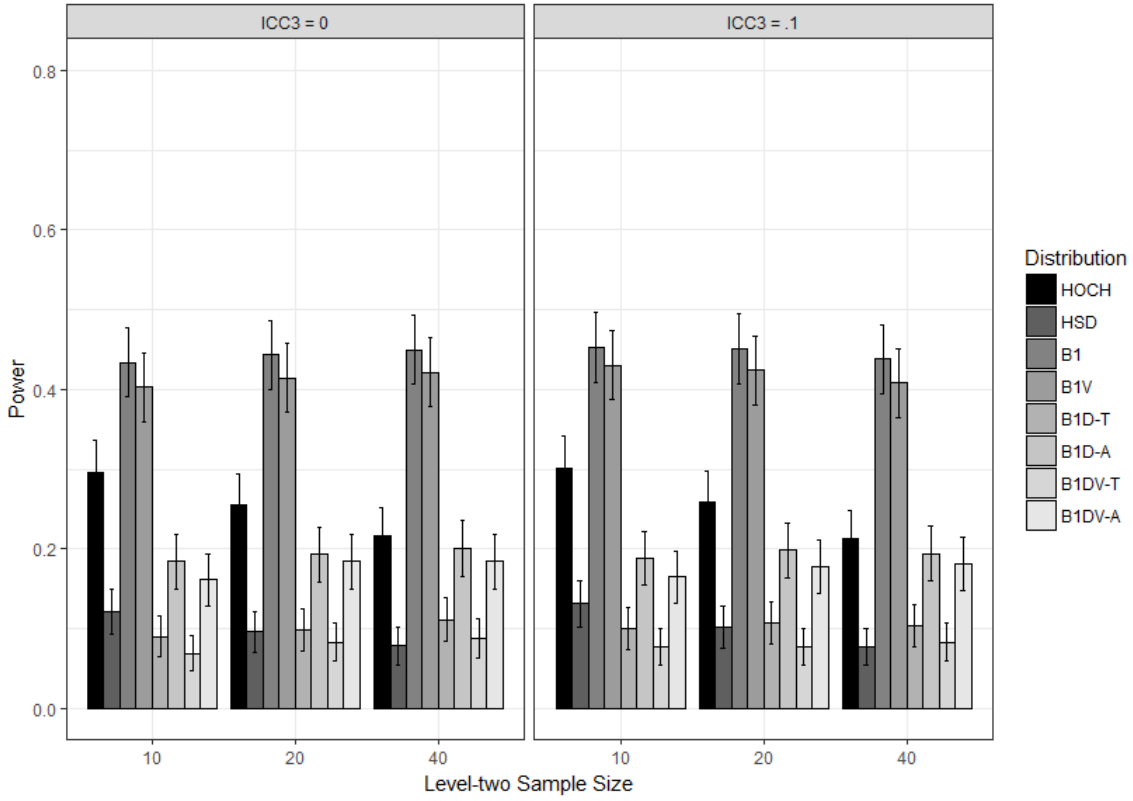


Figure 11. *Power by ICC3 (Homogenous Level-One Variances; Effect = .2; ICC2 = .25)*

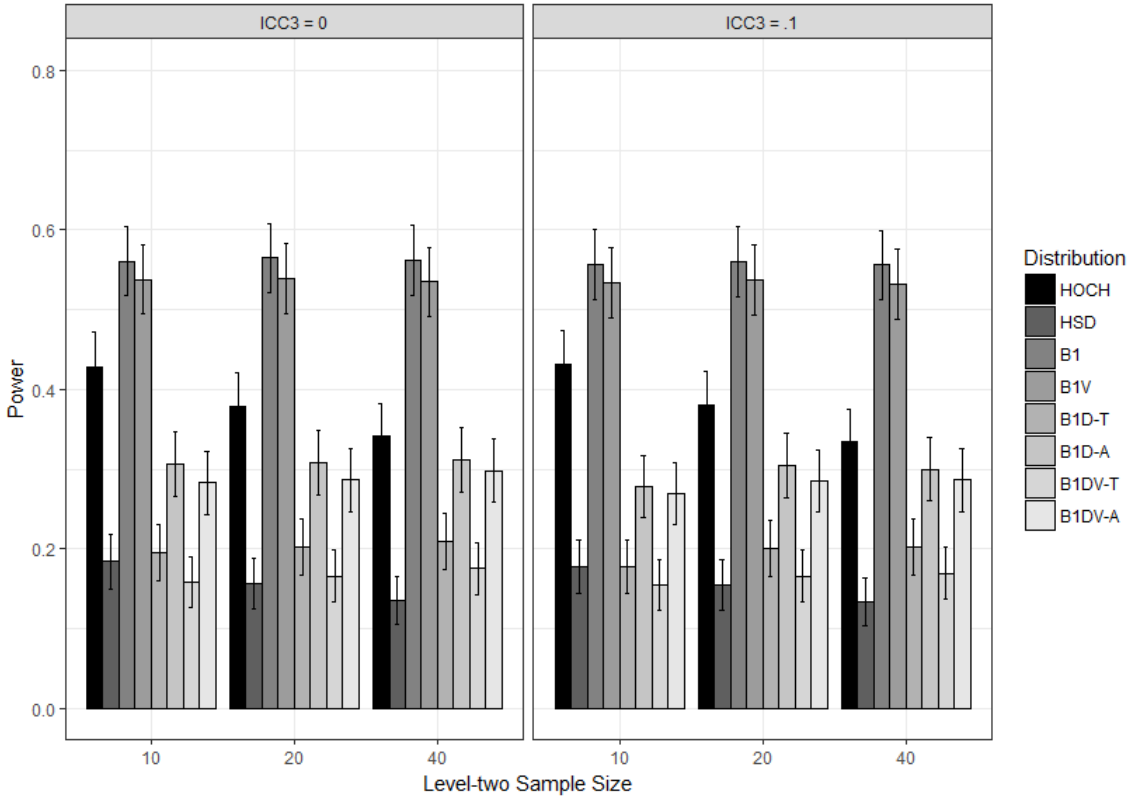


Figure 12. *Power by ICC3 (Homogenous Level-One Variances; Effect = .5; ICC2 = .15)*

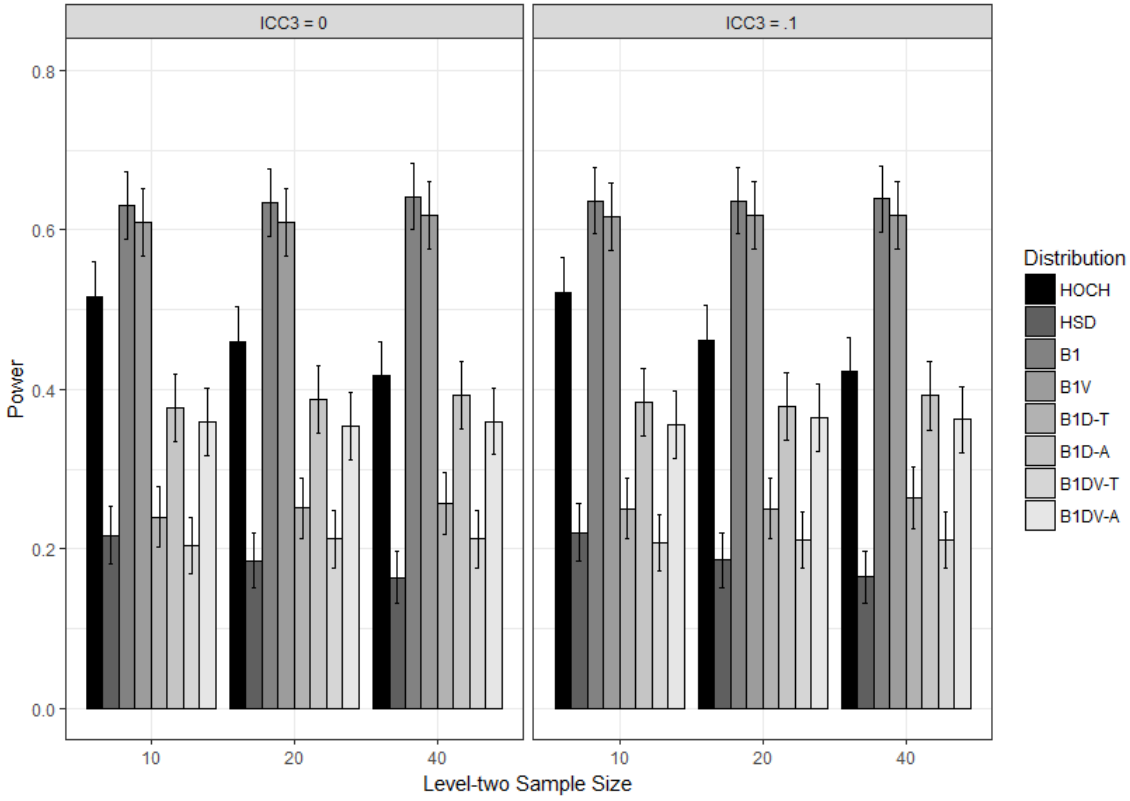
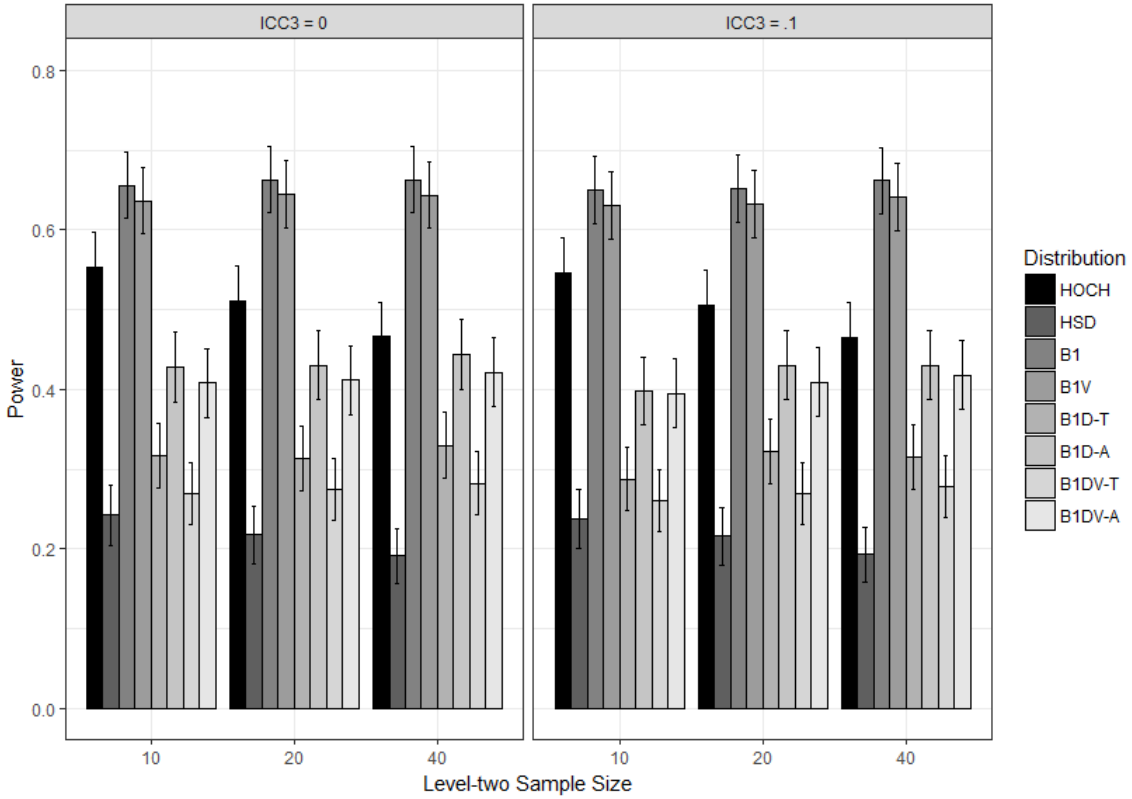


Figure 13. *Power by ICC3 (Homogenous Level-One Variances; Effect = .5; ICC2 = .25)*



Level-one variance heterogeneity.

When there was no higher-level variance present, that is, the conditions in which the *ICC2* and *ICC3* were specified to be zero, the presence of heterogeneous level-one variances generally resulted in greater power for each of the procedures as compared to the condition in which the level-one variances were homogenous. This pattern tended to hold over all levels of the effect size and *N*.

When the *ICC2* was specified to be non-zero, but the *ICC3* was held at zero, the HOCH, B1D and B1DV procedure were more powerful in the heterogeneous level-one variance condition than in the homogenous level-one variance condition. The power of the B1 and B1V procedures was generally unchanged across the two conditions. The

HSD procedure was actually less powerful in the heterogeneous level-one variance condition than in the homogenous level-one variance condition.

When the *ICC3* was specified to be non-zero, the power of the HOCH procedure increased in the heterogeneous level-one variance condition than in the homogenous level-one variance condition. The power of the remaining procedures was generally unchanged or it was not possible to determine a consistent effect of varying the level-one variance homogeneity.

Figure 14. *Power by Level-One Variance Condition (Effect = .2; ICC2 = 0; ICC3 = 0)*

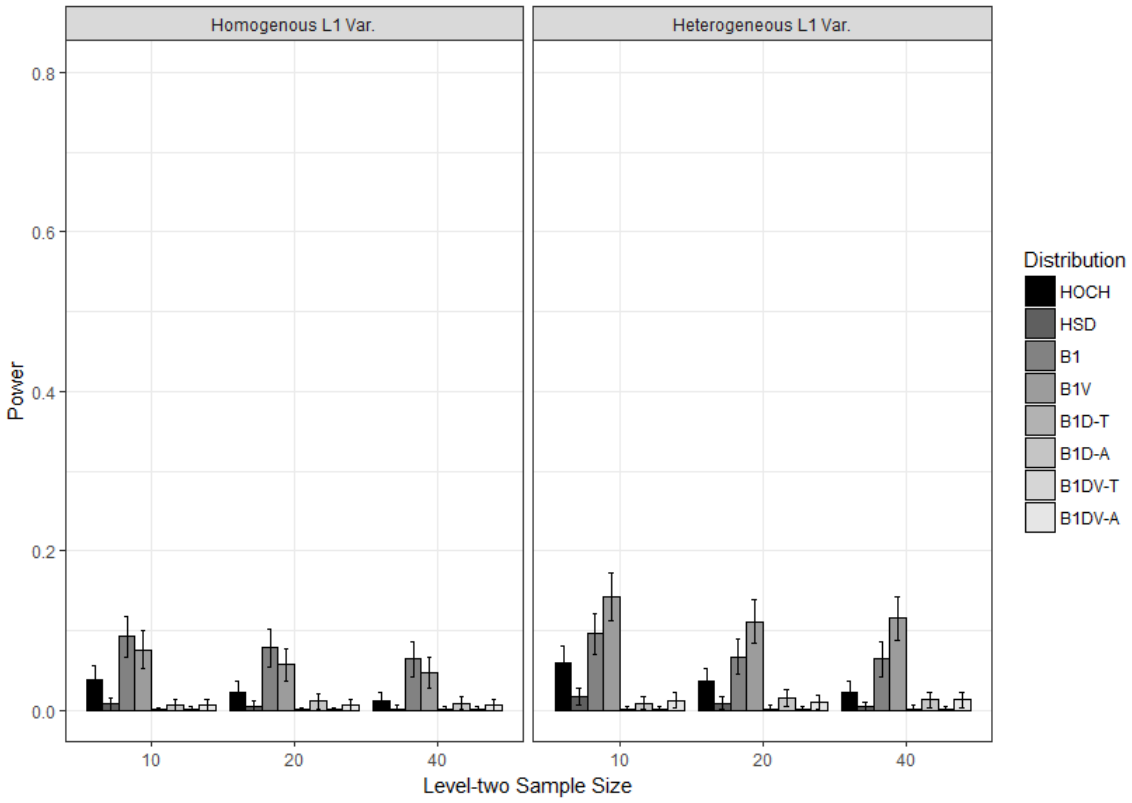
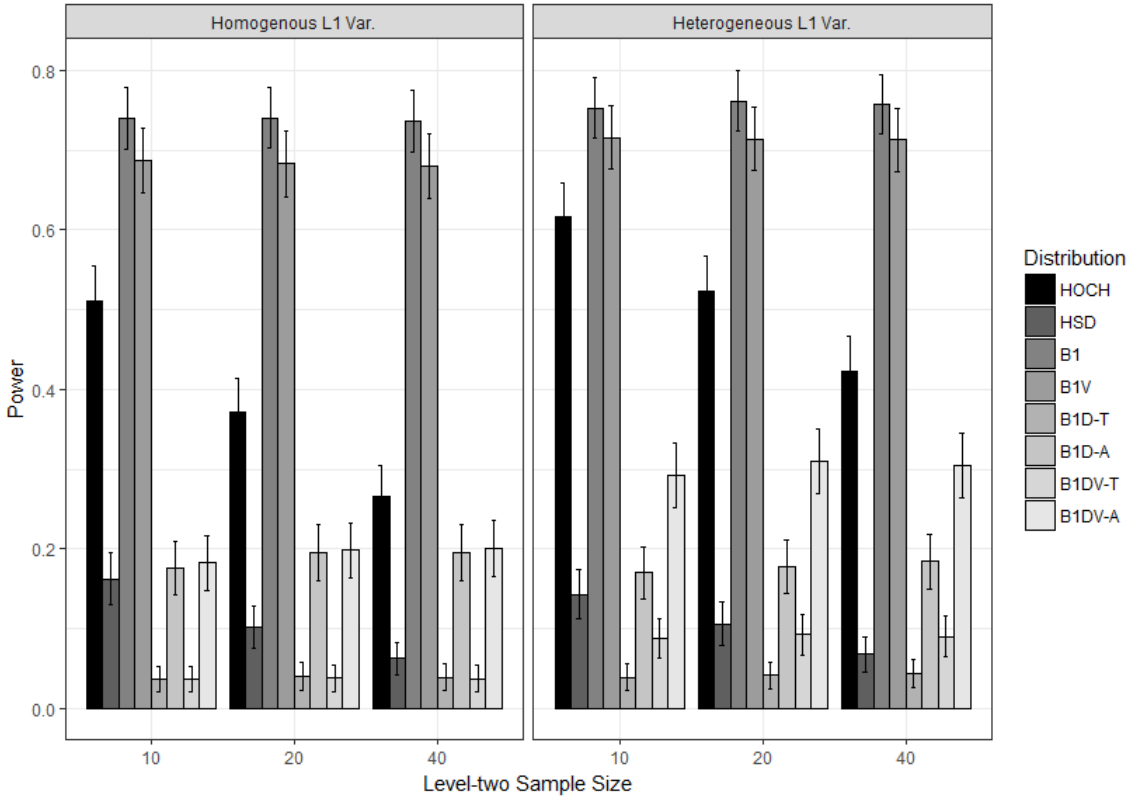


Figure 15. *Power by Level-One Variance Condition (Effect = .5; ICC2 = 0; ICC3 = 0)*



Comparison of the procedures.

As expected, each of the procedures produced a lower average power, across all conditions, when compared to the unadjusted power rates. Across all procedures, the B1 method was the most powerful followed by the B1V and HOCH procedures. The HSD, B1D, and B1DV procedures reported noticeably lower power values as compared to the other three methods. In fact, when β_{ijk} was .2 and no higher-level variance was present, these three procedures had average power values close to zero. The semi informed variance procedures exhibited slightly lower power than did their uninformed counterparts (i.e. the B1V procedure had consistently less power than did the B1 procedure). For the B1D and B1DV procedures, the alternative definition of power produced larger values than did the traditional definition of power.

CHAPTER V. DISCUSSION

The purpose of this study was to compare the performance of four Bayesian models to traditional MCPs in situations where Type I error inflation occurs. Performance was defined as the ability to maintain the Type I error rate below α and as the power to correctly reject the null hypothesis. The study aimed to answer two research questions:

1. When a large number of hypotheses are tested simultaneously, are the Bayesian MLMs able to control the Type I error rate below α while demonstrating greater power than the traditional MCPs?
2. When level-one variance heterogeneity is present, are the Bayesian MLMs able to control the Type I error rate below α while demonstrating greater power than the traditional MCPs?

A Monte Carlo simulation study was performed to provide answers to these research questions. Chapter Five is ordered as follows. First, the conclusions of the simulation study in regards to the above research questions are presented. A general recommendation about the performance of the six procedures – Hochberg's (HOCH), Tukey's HSD (HSD), Bayesian one-way *ANOVA* (B1), Bayesian one-way *ANOVA* with semi-informed variance priors (B1V), Bayesian one-way *ANOVA* with a mean difference parameter (B1D), and Bayesian one-way *ANOVA* with a mean difference parameter and semi-informed variance priors (B1DV) - is then provided. Following these sections, limitations and future directions of the study are discussed. The chapter concludes with a general summary of the study.

Main Findings

Research question 1.

Traditional MCPs are designed to ensure control of the Type I error rate at or below α particularly when a large number of hypotheses are tested simultaneously. The tradeoff for such control of the Type I error rate is that traditional MCPs become more conservative as the number of tested hypotheses increase, resulting in less power to correctly detect false null hypotheses (Kromrey & La Rocca, 1995; Olejnik et al., 1997; Seaman et al., 1991).

Assuming that no higher-level variance is present, the results of the study indicate that the traditional procedures were generally able to maintain the Type I error rate below α for the larger values of N. However, as expected, the power of the HOCH and HSD procedure decreased as the level-two sample size grew larger. The HSD procedure's power was always less than that of the HOCH procedure. For example, when the effect size was medium and the level-one variances were specified to be equal to one another (the homogenous variances condition), the power of the HOCH procedure decreased from 51.1% to 26.5% as the level-two sample size increased from 10 to 40. Likewise, the power of the HSD procedure decreased from 16.2% to 6.2% as the level-two sample size increased from 10 to 40.

The B1 and B1V did not exhibit this problem. Not only was the power of the B1 and B1V procedures larger than the traditional methods across all simulation conditions, but also increasing the level-two sample size did not have a noticeable impact on the power of these procedures in the medium effect size condition. For example, in the medium effect condition and with homogenous level-one variances, the power of the B1

procedure was 74% when *N* was 10 and 73.7% when *N* was 40; the power of the B1V procedure was 68.8% when *N* was 10 and 68% wen *N* was 40. When the effect size was small, and particularly when the level-one variances were homogenous, the power of the B1 and B1V procedures decreased with *N*. However, both procedures were still more powerful than the traditional methods.

In exchange for achieving greater power compared to the traditional MCPs, these two methods generally sacrificed the ability to maintain the Type I error rate at α. While the Type I error rate was far less than the unadjusted Type I error rate, the B1 and B1V procedures may not be appropriate for the applied researcher who wishes to maintain strict control of the Type I error rate at α. This sacrifice was pronounced for the B1 procedure than for the B1V procedure; the B1 procedure only produced a Type I error rate less than .05 when *N* was 40 and the level-one variances were equal to one another. The B1V procedure was able to maintain the Type I error rate below α for *N* as low as 20 when the level-one variances were homogenous and at $N = 40$ when the level-one variances were heterogeneous.

Unfortunately, the B1D and B1DV were much too conservative to be of use as a method for controlling for multiplicity. While these two procedures rarely committed a Type I error, they also displayed a correspondingly low ability to identify false null hypotheses. When testing hypotheses by evaluating the posterior distributions of the level-two means, the B1D procedures produced power rates that were lower than the traditional MCPs. When testing hypotheses by evaluating the posterior distribution of the mean difference parameter, δ_q, the B1D and B1DV procedures returned power rates that were comparable to the HSD procedure. Most likely the reason for the low power of

these two procedures is tied to the prior distribution assigned to the hyperparameter p_q. This hyperparameter determines the probability that δ_q is assigned a point mass prior distribution entirely at zero or a normal prior distribution. When δ_q is assigned a point mass prior distribution entirely at zero the null hypothesis will be retained and when δ_q is assigned a normal distribution as a prior there is a non-zero probability that the null hypothesis will be rejected. The prior distribution assigned to p_q, drawn from a previous study by Li and Shang (2016), may have resulted in an overly conservative model that favored assigning δ_q the point mass prior distribution at zero.

When testing a large number of hypotheses, the B1 procedure is recommended due to its superior power to the other methods and its improved control of the Type I error rate as the number of hypotheses increases. If strict control of the Type I error rate is desired or a small number of hypotheses are being tested, the HOCH procedure is recommended. The above recommendation was formed in the conditions in which no level-two or three variance was present. The above patterns of results hold for the simulation conditions in which the *ICC2* and *ICC3* were specified to be non-zero.

Research question 2.

Variance heterogeneity among the level-one units did not affect the traditional MCP's ability to maintain the Type I error rate below α. Although previous research (Games & Howell, 1976; Hsiung & Olejnik, 1994; Kromrey & La Rocca, 1995) had shown that the power of the traditional methods should have decreased in the presence of level-one variance heterogeneity that was not the case in this study. In fact, both the HOCH and HSD procedures were more powerful when the level-one variances were

heterogeneous than when the level-one variances were specified to be equal to one another.

This result was likely due to the research design used. Previous studies that found evidence of decreased power due to level-one variance heterogeneity tested hypotheses evaluating pairwise differences between level-two means. In these studies, variance heterogeneity increased the standard error of pairwise mean difference test statistics making it more difficult to reject a false null hypothesis. On the other hand, the present study tested hypotheses regarding differences between the level-two means and a constant criterion of the grand centered aggregate mean. Recall, that in the homogenous level-one variance condition, each variance was specified to be one. In the heterogeneous level-one variance condition, half the level-one variances were specified to be .5 and the other half was specified to be 1.5. Not only does variance heterogeneity not necessarily increase the standard error of the test statistic, but those level-two groups that are assigned the smaller variance it the heterogeneous condition are more likely to be correctly flagged as being significantly different than the criterion as compared to the level-two groups in the homogenous level-one variance condition.

Likewise, the Bayesian methods were more powerful when the level-one variances were heterogeneous than when the level-one variances were homogenous. Additionally, when the effect was small the Bayesian methods with adaptive prior distributions on the level-one variances (the B1V and B1DV methods) were more powerful than their non-adaptive counterparts (the B1 and B1D methods) when level-one variance heterogeneity was present. The B1DV method was more powerful than the B1D method in the medium effect condition under variance heterogeneity.

Unfortunately, the Type I error control of the B1 and B1V methods was adversely affected by the presence of level-one variance heterogeneity. Both methods largely failed to maintain the Type I error rate below α when variance heterogeneity was present. The only exception was the B1V method when N was equal to 40. Because the B1D and B1DV methods are overly conservative procedures (as discussed above) they did not encounter this issue when variance heterogeneity was present.

In conclusion, if the researcher is first concerned with maintaining the Type I error rate below α, with power being a secondary concern, then the HOCH procedure is recommended when level-one variance heterogeneity is present. If the researcher is able to be less strict about maintaining the Type I error rate below α, than the B1V procedure is recommended when variance heterogeneity is present due to its ability to correctly detect small effects and increased power as a larger number of hypotheses are tested. The discussion thus far has focused on the conditions in which no level-two or three variance occurred. The above patterns of results hold for the simulation conditions in which the ICC2 and ICC3 were specified to be non-zero.

Overall performance.

Assuming there was no variance present at level-two or three, the HOCH, HSD, B1D, and B1DV procedures were able to maintain the Type I error rate below α. While a Type I error rate in excess of α was reported for some conditions of N for the HOCH procedure this was determined to be due to the result of random simulation error during the data generation process. Data were generated by randomly sampling a normal distribution with a mean of zero. For the simulation seed used in this study, a large enough number of replications contained values taken from the tails of the generating

distribution so that the Type I error rate was greater than alpha. While the probability of this occurring was low, these were legitimate simulation results. To confirm that these results were due to random error, the simulation seed was varied and none of the resulting simulated data sets reproduced a Type I error rate in excess of α for the HOCH procedure. The B1 and B1V struggled to maintain the Type I error rate below α when the level-two sample size was small and when level-one variance heterogeneity was present.

Across all simulation conditions, the B1 and B1V methods demonstrated the most power in correctly rejecting the null hypothesis. Outside of those procedures, the HOCH procedure was the next most powerful followed by the HSD procedure and then the B1D and B1DV procedures. Generally, the Bayesian methods with adaptive prior distributions on the level-one variances were less powerful than their non-adaptive counterparts. This may be the case because the models with adaptive prior distributions over fit the data by providing separate estimates of the level-one variances when a single estimate would have sufficed. The difference in power between these two models did shrink when variance heterogeneity was present. The one exception to this pattern occurred when the level-one variances were heterogeneous and the effect size was small. In this scenario, the B1V model was more powerful than the B1 model. Allowing separate parameter estimates for each level-one variance provided additional information about the grand mean, which, in turn, gave the B1V model more power to correctly reject the null hypothesis as compared to the B1 model.

In conclusion, if the goal of the researcher is to maintain the Type I error rate below α while retaining the greatest power to correctly reject null hypotheses, the HOCH procedure would be preferred if the number of hypotheses being test is less than 40.

Provided the level-two and three variance was zero, the HOCH procedure generally maintained good control of the Type I error rate while being more powerful than the HSD, B1D and B1DV procedures. If the number of hypotheses being tested is large ($N \geq$ 40) then the B1 procedure will maintain control of the Type I error rate below α. As a result, the B1 procedure should be selected over the HOCH procedure due to its greater power in this scenario.

If the researcher is able to accept a liberal Type I error rate, then the B1 or B1V procedures should be chosen. These procedures demonstrated greater power than the other four methods across all simulation conditions. The preference of procedure should be given to the B1 method over the B1V method with the exception of the scenario in which a small number of hypotheses are being tested or the scenario in which the effect is presumed to be small

Under the present simulation conditions, the HSD, B1D, and B1DV procedures cannot be recommended. While these procedures always maintained the Type I error rate below α, these procedures lacked the ability to detect false null hypotheses. A more powerful procedure exists for every condition in which strict control of the Type I error rate would be necessary.

Limitations

The present study had several limitations – some resulting from the conditions in which the simulation took place and others resulting from a lack of resources. Every study is limited by the settings that were not considered. These un-realized settings constrain the generalizability of the study.

One limitation was the research scenario in which this study was set. This study was conducted under the scenario of evaluating several level-two means against a single criterion. An example of this might be evaluating the standardized test scores of all the high schools in a district against the average score for schools across the country. This resulted in several limitations. First, this limits the generalizability of the study to this specific scenario. Additionally, and as explored above, this study design may have confounded the effects of level-one variance heterogeneity on the performance of the methods under consideration. The reason for this is the test statistic that underlies the traditional MCPs in these scenarios. When making pairwise comparisons, the independent samples *t*-test is used while the single samples *t*-test is used when evaluating several groups against a criterion. There is ample evidence that variance heterogeneity negatively affects the power of the independent samples *t*-test (Games & Howell, 1976; Kromrey & La Rocca, 1995; Shaffer, 1995) but this is not necessarily true of the single sample *t*-test. Consider the heterogeneous variance condition in which half the level-two units were to have within groups variances of .5 and the other half being assigned variances of 1.5. It will be easier to correctly reject the null hypothesis for those level-two units with variances of .5 than those with variances of 1.5. Consequently, the power of the traditional MCPs will be inflated when variance heterogeneity is present rather than being decreased as would be expected if one were testing pairwise comparisons. Finally, Tukey's HSD procedure was designed for the research scenario in which every level-two mean is evaluated against one another. In the present study, Tukey's HSD was applied to the scenario in which every level-two mean is evaluated against zero. Practically, this means that, instead of comparing the *q* critical value to the *t*-values taken

from all possible independent samples *t*-tests, the *q* critical value was evaluated against *t*-values taken from all possible single sample *t*-tests. This may have affected the power of that procedure.

A second limitation is the choice of definition for the Type I error rate and power. As discussed in Chapter One, there are different definitions of the Type I error rate and power and the decision of which definition to use in practice is largely influenced by the research design of the study. In this study the familywise Type I error rate and the all-pair power were the chosen definitions. As a result, the generalizability of this study is again limited to only the situations in which the familywise and all-pair power definitions are used.

A further limitation is that this study only considered the scenario in which the level-two units were balanced with respect to the level-one sample size. Consequently, the results of this study may only be generalized to the situations in which all level-one sample sizes are equal.

The study is also limited by the values and distributions chosen as the prior distributions for the Bayesian models. An infinite combination of prior distributions can be chosen for the parameter and hyperparameters for a given models and discussion of alternative prior distributions is generally outside the scope of this paper. One exception to that is the prior distribution assigned to p_q in the B1D and B1DV model. This parameter determines the probability that the prior distribution on δ_q is either zero or a normal distribution. Previous studies (Li & Shang, 2016) have assigned this parameter a *BETA* distribution with the suggestion that $PR\{\delta_i = 0\} = .5 = \frac{A_0}{A_0 + B_0}$ for the

hyperparamters A_0 and B_0. This suggestion resulted in overly conservative models and other values for this equality should be considered.

Additonally, the posterior distributions of the parameters of B1D and B1DV models had difficulty meeting the convergence criterion of a $\hat{R}$ value less than 1.1. There are two main ways to assist a model in achieving convergence. The first would be to assign the prior distributions in the model more informative prior distributions. This was not done because the study was designed to evaluate the performance of these procedures using noninformative prior distributions. The second method would be to use brute force to increase the likelihood of achieving model convergence by increasing the number of draws from the posterior distribution. As a result, the B1D and B1DV models were evaluated when they met the more lenient criterion of producing a $\hat{R}$ value less than 1.5.

Future Directions

Future research is needed to determine the generalizability of the use Bayesian models as MCPs outside the settings considered in the present study. Much like the limitations of this study, there are a myriad of directions in which research on this topic could be extended. In the discussion below, possibilities for incorporating a few of the more salient extensions into future research are presented.

Future studies should investigate the extent to which different prior distributions of the parameters of the Bayesian models affect their ability to perform as MCPs. Specifically, researchers should focus on the prior distributions on the mean parameters and, for the B1D and B1DV models, on δ_q and its hyperparameters. This could be done in two ways. First, this could be accomplished by exploring the effect of different uninformative prior distributions. For example, in the current study the mean parameters

were assigned normal prior distributions. Future studies could investigate the performance of these models when the mean parameters are assigned a uniform prior distribution. Second, researchers could investigate scenarios in which it may be appropriate to specify informative prior distributions. For instance, if the rank order assumption is tenable, as discussed in Chapter One, it may be reasonable to specify prior distributions that take into account the expected ordering or the mean parameters.

Along the same lines, future research could focus on the variables used to achieve convergence of the Bayesian models. There are a variety of ways researchers could explore to speed up the process of convergence. Researchers could increase the number of MCMC chains, differ the amount of thinning of the MCMC chains or increase the number of draws taken from the posterior distribution. Additionally, other criterion for evaluating convergence outside the $\hat{R}$ statistic could be considered. One interesting extension would be to vary the software used to estimate the Bayesian models.

Another avenue for research would be to expand the research scenarios in which the performance of these methods are evaluated. Many traditional MCPs were developed for research scenarios not included in this study. Specifically, Tukey's HSD was developed for the scenario in which one wishes to make all pairwise comparisons among means (Toothaker, 1991; Tukey, 1953) and Dunnett's test was developed for the scenario in which one wishes to evaluate every mean versus a single control mean (Dunnett, 1955). Future research could evaluate the performance of the Bayesian methods against traditional MCPs which were developed with these different research designs in mind.

Of particular interest to future researchers might be the false discovery rate. Traditional MCPs that are designed to control for the familywise Type I error rate at α are

known to decrease in power as the number of tested hypotheses tested increase (Kromrey & La Rocca, 1995; Olejnik et al., 1997; Seaman et al., 1991) and more evidence of that phenomenon was shown in the present study. Consequently, one of the motivations behind this study was to evaluate the power of the procedures in this setting. The false discovery rate has been more commonly used in scenarios in which a large number of hypotheses are tested simultaneously (cite) and MCPs that control for the false discovery rate have been found to be more powerful than procedures that control for the familywise error rate. There is evidence that methods that control for the false discovery rate are able to better handle a larger number of tests than methods that control for the familywise error rate (Lu & Westfall, 2009; Westfall, 2010). The Bayesian models investigated in this study could be evaluated in terms of false discovery rate controlling procedures.

Conclusions

This study provides several contributions to the literature on multiple comparisons procedures. First, it is the one of the first empirical evaluations of the ability of Bayesian models to act as a control for Type I error inflation, particularly in comparison to traditional MCPs. Second, this study provided evidence of how these Bayesian models perform when a large number of hypotheses are tested simultaneously and when variance heterogeneity is present – two scenarios that have been shown to be detrimental to the performance of traditional MCPs. Finally, this study gave evidence about how different conceptualizations of Bayesian models, through either adapting the prior distributions to account for variance heterogeneity or by reformulating how the models conceived mean differences, affected the Type I error control and power of these methods.

The results of this study have implications for the applied researcher. This study provided evidence that, in the scenario, when over 40 hypotheses are being tested simultaneously the Bayesian one-way *ANOVA* should be preferred to traditional MCPs due to the model's control of the Type I error rate and high power. In scenarios in which a smaller number of hypotheses are tested, the Bayesian models cannot be recommended over the traditional MCP. Finally, this study provided negligible evidence that allowing the prior distributions on the level-one variances to differ had any impact on the performance of these models. Consequently, it is recommended that if researchers decided to utilize the Bayesian one-way *ANOVA* as a control for multiplicity that the simpler model with a single prior distribution on the level-one variance be chosen.

References

Abdi, H. (2010). Holm's Sequential Bonferroni Procedure. In N.J. Salkind (Ed.), *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage.

Aikin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferrroni vs. Holm methods. *American Journal of Public Health, 86*, 726-728.

Austin, P.C., Naylor, C., & Tu, J.V. (2001). A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *Journal of Evaluation in Clinical Practice, 7*, 35-45.

Austin, P.C. (2010). Estimating multilevel logistic regression models when the number of clusters is low; a comparison of different statistical software procedures. *The International Journal of Biostatistics, 6,* 1-18.

Baldwin, S.A., & Fellingham, G.W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*, 151-164.

Bayarri, M.J., & Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19*, 58-80.

Bender, R., & Lange, S. (2001). Adjusting for multiple testing – when and how? *Journal of Clinical Epidemiology, 54*, 343-349.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300.

Berry, D.A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference, 82*, 215-227.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8,* 3-62.

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple Comparisons using R*. Boca Raton, FL: Taylor & Francis Group.

Brown, B.W., & Russell, K. (1997). Methods correcting for multiple testing: Operating characteristics. *Statistics in Medicine, 16*, 2511-2528.

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

Cunningham, T. (2010). *Power and sample size for three-level cluster designs (Unpublished doctoral dissertation).* Virginia Commonwealth University, Richmond, VA.

deCani, J.S. (1984). Balancing type I risk and loss of power in ordered Bonferroni procedures. *Journal of Educational Psychology, 76*, 1035-1037.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London, England: Arnold.

Donoghue, J. R. (1998). Implementing Shaffer's Multiple Comparison Procedure for Large Numbers of Groups. *ETS Research Report Series, 1998*(2), i-38.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*, 52-64.

Dunnett C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*, 1096–1121.

Dunnet, C.W., & Tamhane, A.C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association, 87*, 162-170.

Einot, I., & Gabriel, K.R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association, 70*, 574-583.

Fazzari, M. J., Kim, M. Y., & Heo, M. (2014). Sample size determination for three-level randomized clinical trials with randomization at the first or second level. *Journal of Biopharmaceutical Statistics, 24*, 579-599.

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness, 6*, 44-57.

Freedman, L. (1996). Bayesian statistical methods: A natural way to assess clinical evidence. *British Medical Journal, 313*, 569-570.

Games, P.A. & Howell, J.F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances. *Journal of Educational Statistics, 1*, 113-125.

Games, P.A., Keselman, H.J., & Rogan, J.C. (1981). Simultaneous pairwise multiple comparison procedures for means when sample sizes are unequal. *Psychological Bulletin, 90*, 594-598.

Gelman, A. (1996). Inference and Monitoring Convergence. In W.R. Gilks, S. Richardson & D.J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in Practice*. London, England: Chapman and Hall.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 3*, 515-534.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 189- 211.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of college performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159*, 149-163.

Grandes, G., Sanchez, A., Sanchez-Pinilla R. O., Torcal J., Montoya, I., Lizarraga, K., & Serra, J. (2009). Effectiveness of physical activity advice and prescription by physicians in routine primary care. *Archives of Internal Medicine, 7*, 694-701.

Gravetter, F.J., & Wallnau, L.B. (2017). *Essentials of Statistics for the Behavioral Sciences* (10th Ed.). Belmont, CA: Thompson Higher Education.

Greenland, S. (2000). Principles of multilevel modeling. *International Journal of Epidemiology, 29*, 158-167.

Hays, W.L. (1994). *Statistics* (5th Ed.). Belmont, CA: Wadsworth Group/Thompson Learning.

Hedges, L. V., & Hedberg, E. C. (2007a). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education, 22*, 1-15.

Hedges, L. V., & Hedberg, E. C. (2007b). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-802.

Hochberg, Y., & Tamhane, A.C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Taylor & Francis.

Holland, B.S., & Copenhaver, M.D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin, 104*, 145-149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*, 383-386.

Horn, M., & Dunnett, C.W. (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. *Lecture Notes-Monograph Series, 47*, 48-64.

Hox, J., van de Schoot, R., & Matthijse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6,* 87-93.

Hsiung, T. H., & Olejnik, S. (1994). Power of pairwise multiple comparisons in the unequal variance case. *Communications in Statistics-Simulation and Computation, 23,* 691-710.

Jones, H.E., Ohlssen, D.I., Neuenschwander, B., Racine, A., & Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials, 8,* 129-143.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY: The Guilford Press.

Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook* (4th Ed.). Upper Saddle River, NJ: Pearson Education.

Kim, S., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A Review. *Journal of Educational and Behavioral Statistics, 23,* 356-377.

Klockars, A. J., & Hancock, G.R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychological Bulletin, 11,* 505-510.

Kreft, I.G.G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University, Los Angeles, CA.

Kromrey, J.D., & La Rocca, M.A. (1995). Power and type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. *The Journal of Experimental Education, 63*, 343-362.

Kruschke, J.K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology*, 142, 573-603.

Kruschke, J.K, & Meredith, M. (2015). BEST: Bayesian estimation supersedes the t test. R package version 0.40. http://CRAN.R-project.org/package=BEST

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*, 835-851.

Lehmann, E.L., & Romano, J.P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics, 33*, 1138-1154.

Levin, B. (1996). Annotation: On the Holm, Simes, and Hochberg multiple test procedures. *American Journal of Public Health, 86*, 628-629.

Li, Q., & Shang, J. (2015). A Bayesian hierarchical model for multiple comparisons and mixedmodels. *Communications in Statistics – Theory and Methods, 44*, 5701-5090.

Lu, Y., & Westfall, P.H. (2009). Is Bonferroni admissible for large *m*? *American Journal of Mathematical and Management Sciences, 29*, 51-69.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology, 25*, 1032-1037.

Ludbrook, J. (1991). On making multiple comparisons in clinical and experimental pharmacology and physiology. *Clinical and Experimental Pharmacology and Physiology, 18*, 379-392.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*, 3049-3067.

Lynch, S.M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Princeton, NJ: Springer.

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

Martin, A.D., Quinn, K.M., & Park, J.H. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software, 42*(9), 1-21.

McNeish, D.M., & Stapleton, L.M. (2016). The effect of small sample size of two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295-214.

Metropolis, N., Rosenbluth, A. W., Rosenluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 212*, 1087-1091.

Mossman, D., & Berger, J. (2001). Intervals for post-test probabilities: A comparison of five methods. *Medical Decision Making, 21*, 498-507.

Murray, D.M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Nashimoto, K., & Wright, F.T. (2008). Bayesian multiple comparisons of simply ordered means using priors with a point mass. *Computational Statistics and Data Analysis, 52*, 5143-5153.

National Center for Education Statistics (1997). *1996 NAEP comparisons of average scores for participating jurisdictions*. Washington, D.C.: Government Printing Office.

Nee, D. (2014, December 12). Multiple statistical tests. Retrieved from http://danielnee.com/tag/hommel.

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 et seq. (West 2003)

Normand, S.T., Glickman, M.E., & Gastonis, C.A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, 92, 803-814.

Ohlessen, D.I., Sharples, L.D., & Spiegelhalter, D.J. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine, 26*, 2088-2112.

O'Keefe, D.J. (2003). Colloquy: Should familywise alpha be adjusted? *Human Communication Research, 29*, 431-447.

Olejnik, S., Li, J., Supattathum, & Huberty, C.J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics, 22*, 389-406.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and Prediction* (3rd ed.). Belmont, CA: Thompson Learning, Inc.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. Retrieved from http://www.R-project.org.

Ramsey, P.H. (1981). Power of univariate pairwise multiple comparison procedures. *Psychological Bulletin, 90,* 352-366.

Ramsey, P.H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association, 73,* 479-485.

Ramsey, P.H., Ramsey, P.P., & Barrera, K. (2010). Choosing the best pairwise comparison of means from non-normal populations, with unequal variances, but equal sample sizes. *Journal of Statistical Computation and Simulation, 80,* 595-608.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13,* 85-116.

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

Raudenbush, S.W., & Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307-335.

Ryan, T.A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56,* 26-47.

Scheffé, H. (1970). Practical solutions to the Behrens-Fisher problem. *Journal of the American Statistical Association, 65,* 1501-1508.

Scheffé, H. (1959). *The Analysis of Variance*. New York, NY: John Wiley.

Schochet, P.Z., & Chiang, H.S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142-171.

Seaman, M.A., Levin, J.R., & Serlin, R.C. (1991). New developments in pairwise multiple comparison: Some powerful and practicable procedures. *Psychological Bulletin, 110*, 577-586.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology, 46*, 561-584.

Shang, J. (2011). A Bayesian multiple comparison procedure for simple order-restricted mixed models with missing values. *Journal of Data Science, 9*, 311-330.

Shang, J., Cavanaugh, J.E., & Wright, F.T. (2008). A Bayesian multiple comparison procedure for order-restricted mixed models. *International Statistical Review, 76*, 268-284.

Shaw, L.H. (2012). *Incorporating latent variable outcomes in value-added assessment: An evaluation of univariate and multivariate model structures* (Unpublished doctoral dissertation). University of Nebraska-Lincoln, Lincoln, NE.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626-633.

Siddiqui, O., Hedeker, D., Flay, B.R., & Hu, F.B. (1996). Intraclass correlation estimates in a school-based smoking prevention study: Outcome and

mediating variables, by sex and ethnicity. *American Journal of Epidemiology, 144*, 425-433.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*, 751-754.

Snijders, T. A. B. (2005). Power and sample size in multilevel modeling. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Chicester, England: Wiley.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.) London, England: Sage.

Strassburger, K., & Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in medicine, 27*, 4914-4927.

Statisticat, LLC. (2016). LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. R package version 16.1.1. [https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software]

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science, 57*, 748-761.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical software, 12*, 1-16.

Tabachnick, B.G. & Fidell. L.S. (2013). *Using Multivariate Statistics* (6th Ed.). Boston, MA: Pearson Education.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271-288.

Toothaker, L.E. (1991). *Multiple comparisons for researchers*. Newbury, CA: Sage.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript.

Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., & Drazen, J.M. (2007). Statistics in medicine – reporting of subgroup analysis in clinical trials. *The New England Journal of Medicine, 357*, 2189-2194.

Wang, Y.G., Leung, D.H.Y., Li, M., & T, S.B. (2005). Bayesian designs with frequentist and Bayesian error rate considerations. *Statistical Methods in Medical Research, 14*, 445-456.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistical Association, 70*, 129-133.

Westfall, P.H., Johnson, W.O., & Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika, 84*, 419-427.

Westfall, P. H., & Wolfinger, R. D. (2000). Closed multiple testing procedures and PROC MULTTEST. *SAS institute Inc*.

Westfall, P.H. (2010). How well do multiple testing methods scale up when both *n* and *k* increase? *Journal of Biopharmaceutical Statistics, 21*, 583-594.

Wright, S.P. (1992). Adjusted *p*-values for simultaneous inference. *Biometrics, 48*, 1005-1013.

Yi, N., Xu, S., Lou, X.Y., & Mallick, H. (2014). Multiple comparisons in genetic association studies: A hierarchical modeling approach. *Statistical Applications in Genetics and Molecular Biology, 13*, 35-48.

APPENDIX A: PROPERTIES OF MCPS FOR PAIRWISE COMPARISONS

Many of the traditional MCPs were designed with the goal of making pairwise comparisons between means. These procedures have properties associated with them that are not applicable for procedures that make multiple comparisons against a criterion. Below is a discussion of several of these properties.

Closure

A set of hypotheses are said to be closed if the set contains all original hypotheses along with all hypotheses that are formed by the interaction of the original hypotheses (Shaffer, 1995). This is most easily explained in the situation in which one is making pairwise comparisons. To demonstrate what constitutes a closed set of hypotheses, assume that three means are to be compared. H_{12} reflects the hypothesis that tests whether the population means for group 1 and 2 are equal. Likewise, H_{123} would test the hypothesis that $\mu_1 = \mu_2 = \mu_3$. When testing all pairwise comparisons among three groups, the relevant set of hypotheses is H_{12}, H_{13}, and H_{23}. The intersection of a set of hypotheses is all hypotheses formed by the inclusion of the original hypothesis. In the above pairwise comparison set of hypotheses, the intersection would be H_{123} or $\mu_1 = \mu_2 = \mu_3$. H_{123} is also said to be above hypotheses H_{12}, H_{13}, and H_{23} in the hierarchy of hypotheses. The hypotheses that form the intersection are referred to as proper components. If the null hypothesis is rejected for a bivariate comparison of means, it is inappropriate to retain the null hypothesis for the intersection of those hypotheses.

The closure of a set of hypotheses occurs if a hypothesis is rejected at α and every hypothesis that occurs above it in the hierarchy of hypotheses is rejected as well (Shaffer, 1995; Westfall & Wolfinger, 2000). This principle, also known as *coherence*, *consonance* or the *property of free combination* (Holm, 1979; Levin, 1996; Wright, 1992), is a characteristic of most MCPs (Einot & Gabriel, 1975). Many MCPs are designed to be coherent by analyzing hypotheses sequentially (Einot & Gabriel, 1975). *Dissonance* occurs when an intersection of hypotheses is rejected but none of the proper components of the intersection of hypotheses are rejected (Einot & Gabriel, 1975). This is equivalent to declaring an omnibus statistic significant and then finding none of the pairwise, adjusted *p*-values to be significant. MCPs that are formed using closed hypotheses maintain the familywise error rate at α (Shaffer, 1995). MCPs that assure coherence avoid logical contradictions in rejecting hypotheses. In addition, MCPs that test a closed set of hypotheses are guaranteed to maintain strong control of the Type I error rate. These procedures are more powerful than other MCPs that maintain strong control of the Type I error rate but that do not tests a closed set of hypotheses (Shaffer, 1995). The majority of sequential procedures utilize the closure property of hypothesis testing (Westfall & Wolfinger, 2000).

Variance Heterogeneity

The reasons for this power loss with the Tukey's HSD and Scheffé's procedures can be seen by examining the denominator in Equations 14 and 15, which utilizes the *MSE* obtained from an omnibus *ANOVA*. The presence of variance heterogeneity results in larger values of the *MSE* as compared to when

all level two units have equal variance. This in turn decreases the Tukey's HSD and Scheffé's test statistic's magnitude making it more difficult to declare any pairwise comparison significantly different. Likewise, when making pairwise comparisons among level two means, the Bonferroni based MCPs take *p*-values from several independent samples *t*-tests:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{SS_1 + SS_2}{n_1 - n_2 - 2}(\frac{1}{n_1} + \frac{1}{n_2})}}. \quad (62)$$

If variance heterogeneity is present, the denominator in Equation 61 will increase, resulting in a smaller test statistic and a correspondingly higher *p*-value. This corresponds with the well-known Behrens-Fisher problem (Kim & Cohen, 1998; Scheffé, 1970). In practice, especially when employing quasi-experimental and correlational research designs, it may be impossible to ensure the variance homogeneity assumption is satisfied.

Unequal Level One Sample Sizes

On the other hand, level two units with unequal level one sample sizes may affect Type I error control and the power of MCPs in the pairwise comparisons situation (Nashimoto & Wright, 2008). Specifically, Tukey's HSD's ability to control the Type I error rate has been shown to be adversely affected by unequal sample sizes (Games & Howell, 1976; Games, Keselman, & Rogan, 1981). In contrast, Scheffé's procedure is more robust to unequal sample sizes. Bonferroni based procedures are slightly affected when unequal sample sizes are present because the underlying independent samples *t*-test is affected by unequal sample sizes. However, this effect is negligible and if a researcher is truly

concerned with the effect of unequal sample sizes, *p*-values may be drawn from a more robust test such as the Satterthwaite correction to the *t*-test.

Pattern of Mean Differences

In a study where making pairwise comparisons among means is of interest, the distance between means (referred to as the pattern of mean differences) will influence the Type I error control and power of any MCP. For example, a study which has a pattern of mean differences in which each mean differs from one another by a large amount will demonstrate greater power than a study in which all but one of the means are equal to one another and the remaining mean is only slightly greater than the other means. In a study comparing multiple means, a variety of mean configurations may be present (Ramsey et al., 2010). Below are several common patterns of mean differences: The equally spaced null pairs configuration:

$$\mu_{D_{12}} = \mu_{D_{23}} = \mu_{D_{34}}; \mu_{D_{ij}} \neq 0, \tag{63}$$

The equally spaced null pair configuration:

$$\mu_1 = \mu_2 \neq \mu_3 = \mu_4, \tag{64}$$

And the single extreme mean configuration:

$$\mu_1 = \mu_2 = \mu_3 \neq \mu_4. \tag{65}$$

When comparing statistics drawn from the *F* and studentized *q* distributions it is important to consider two other configurations (Ramsey, 1981): (1) The minimum range configuration and (2) the maximum range configuration. The minimum range configuration occurs for an even number of means when:

$$\mu_1 = \ldots = \mu_{N/2} = -\sigma f,$$
$$\mu_{N/2+1} = \ldots = \mu_N = \sigma f. \quad (66)$$

For an odd number of means the minimum range configuration occurs when:

$$\mu_1 = \cdots = \mu_{(N+1)/2} = -[(N-1)/(N+1)]^{1/2}\sigma f,$$
$$\mu_{(N+3)/2} = \cdots = \mu_N = -[(N-1)/(N+1)]^{1/2}\sigma f, \quad (67)$$

and the maximum range configuration spaces the means such that:

$$\mu_1 = -(N/2)^{1/2}\sigma f,$$
$$\mu_2 = \cdots = \mu_{N-1} = 0,$$
$$\mu_3 = (N/2)^{1/2}\sigma f, \quad (68)$$

where *N* is equal to the number of level two units, σ is the homogeneous within group variance, and *f* is equal to the ratio of the between and within group variances (Ramsey, 1981). Research has shown that as the number of non-null hypotheses increase, so does the familywise error rate and the power to detect true differences (Brown & Russell, 1997; Klockars & Hancock, 1992; Olejnik et al., 1997). Additionally, as the distance between means increases, the power of MCPs increases as well (Brown and Russell, 1997; Klockars and Hancock, 1992; Ramsey, 1981; Ramsey et al., 2010; Seaman et al., 1991).

APPENDIX B: SYNTAX

Data Generation

```
library("plyr")
set.seed(1987)

setwd("File Path")
options(scipen=999)

#Factors
l2n=c(10,20,40)
b1=c(0,.2,.5)
icc2=c(0,.15,.25)
icc3=c(0,.1)

numcond = length(l2n) * length(b1)  * length(icc2) * length(icc3)
conditions = matrix(1:numcond,numcond,5)
colnames(conditions) = c("Condition ID", "Level 2 N", "Effect", "icc2", "icc3")
#Level 2 Sample Size
conditions[,2] = rep(1:length(l2n), each = numcond/length(l2n))
#Effect Size
conditions[,3] = rep(1:length(b1), numcond/length(b1))
conditions = conditions[order(conditions[,3]),]
#ICC 2
conditions[,4] = rep(1:length(icc2), numcond/length(icc2))
conditions = conditions[order(conditions[,4]),]
#ICC 3
conditions[,5] = rep(1:length(icc3), numcond/length(icc3))
conditions = conditions[order(conditions[,5]),]
conditions = unique(conditions[,-1])
numcond = nrow(conditions)

reps=500

#Equal level one variance generation
for(condrep in 1:numcond){
 l1n=30
 l3n=5
 gamma000=0
 sigma2e=1
 n2=l2n[conditions[condrep,1]]
 beta=b1[conditions[condrep,2]]
 ICC2=icc2[conditions[condrep,3]]
 ICC3=icc3[conditions[condrep,4]]
 var2=ifelse(ICC2==.25&ICC3==0,.3333,
          ifelse(ICC2==.15&ICC3==0,.17647,
```

```
            ifelse(ICC2==.25&ICC3==.1,.29997,
            ifelse(ICC2==.15&ICC3==.1,.15885,0))))
  var3=ifelse(ICC3==.1&ICC2==0,.1,
            ifelse(ICC3==.1&ICC2==.15,.01765,
            ifelse(ICC3==.1&ICC2==.25,.03333,0)))
  samp.tot=n2*l1n

    for(a in 1:reps){
      group2=rep(1:n2,each=l1n)
      group3=rep(1:l3n,each=(n2/5)*l1n)
      eijk=rnorm(samp.tot,0,sigma2e)
      U0jk=rep(rnorm(n2,0,sqrt(var2)),each=l1n)
      V00k=rep(rnorm(l3n,0, sqrt(var3)),each=l1n*(n2/l3n))
      x=rep(c(-1,1),each=samp.tot/2)
      yijk=gamma000+(beta*x)+eijk+U0jk+V00k
      dat=data.frame(yijk,group2,group3,x,eijk,U0jk,V00k)
      names(dat)=c("y","level2","level3","x","eijk","U0jk","V00k")
      dat$y=dat$y-mean(dat$y)

      write.csv(dat,paste("eqvar.cond",condrep,".rep",a,".csv",sep=""))
    }
}

####################################Unequal level one variance generation
for(condrep in 1:numcond){
  l1n=30
  l3n=5
  gamma000=0
  n2=l2n[conditions[condrep,1]]
  beta=b1[conditions[condrep,2]]
  ICC2=icc2[conditions[condrep,3]]
  ICC3=icc3[conditions[condrep,4]]
  var2=ifelse(ICC2==.25&ICC3==0,.3333,
            ifelse(ICC2==.15&ICC3==0,.17647,
            ifelse(ICC2==.25&ICC3==.1,.29997,
            ifelse(ICC2==.15&ICC3==.1,.15885,0))))
  var3=ifelse(ICC3==.1&ICC2==0,.1,
            ifelse(ICC3==.1&ICC2==.15,.01765,
            ifelse(ICC3==.1&ICC2==.25,.03333,0)))

  samp.tot=n2*l1n

    for(a in 1:reps){
      group2=rep(1:n2,each=l1n)
      group3=rep(1:l3n,each=(n2/5)*l1n)
      eijk=c(rnorm(samp.tot/2,0,sqrt(.5)),rnorm(samp.tot/2,0,sqrt(1.5)))
```

```
    U0jk=rep(rnorm(n2,0,sqrt(var2)),each=l1n)
    V00k=rep(rnorm(l3n,0, sqrt(var3)),each=l1n*(n2/l3n))
    x=rep(c(-1,1),each=samp.tot/2)
    yijk=gamma000+(beta*x)+eijk+U0jk+V00k
    dat=data.frame(yijk,group2,group3,x)
    names(dat)=c("y","level2","level3","x")
    dat$y=dat$y-mean(dat$y)

    write.csv(dat,paste("uneqvar.cond",condrep,".rep",a,".csv",sep=""))
   }
}
```

Traditional MCPS

```
matsig.eq=matpow.eq=matsig.uneq=matpow.uneq=list()

##################Equal Level One Variance Condition
for(condrep in 1:numcond){
 l1n=30
 l3n=5
 alpha=.05

 n2=l2n[conditions[condrep,1]]
 beta=b1[conditions[condrep,2]]
 ICC2=icc2[conditions[condrep,3]]
 ICC3=icc3[conditions[condrep,4]]
 var2=ifelse(ICC2==.25&ICC3==0,.3333,
          ifelse(ICC2==.15&ICC3==0,.17647,
          ifelse(ICC2==.25&ICC3==.1,.29997,
          ifelse(ICC2==.15&ICC3==.1,.15885,0))))
 var3=ifelse(ICC3==.1&ICC2==0,.1,
          ifelse(ICC3==.1&ICC2==.15,.01765,
          ifelse(ICC3==.1&ICC2==.25,.03333,0)))
 samp.tot=l1n*n2

 mat.sig=matrix(nrow=reps,ncol=10)
 mat.pow=matrix(nrow=reps,ncol=7)

 for(a in 1:reps){
  dfw=(l1n*n2)-n2

  dat=read.csv(paste(getwd(),"/eqvar.cond",condrep,".rep",a,".csv",sep=""),header=T)
  dat=dat[,-1]
  names(dat)=c("y","level2","level3","x")

  t.raw=matrix(ncol=2,nrow=n2)
  for(j in 1:n2){
```

```
    t.raw[j,1]=t.test(dat$y[dat$level2==j])$statistic
    t.raw[j,2]=t.test(dat$y[dat$level2==j])$p.value
    colnames(t.raw)=c("t","unadj p")
   }

   p.hoch=as.matrix(p.adjust(t.raw[,2],method="hochberg"))
   qcrit=qtukey(.95,nm=n2,df=dfw,lower.tail=T)/sqrt(2)
    tapply(dat$y,dat$level2,mean)/sqrt(ms.w/l1n)

   sig.raw=ifelse(sum(ifelse(t.raw[,2]<.05,1,0))>0,1,0)
   sig.hoch=ifelse(sum(ifelse(p.hoch<.05,1,0))>0,1,0)
   sig.tuk=ifelse(sum(ifelse(abs(t.raw[,1]>qcrit),1,0))>0,1,0)

   tot.sig=sum(ifelse(t.raw[,2]<.05,1,0))/n2
   tot.sig.hoch=sum(ifelse(p.hoch<.05,1,0))/n2
   tot.sig.tuk=sum(ifelse(abs(t.raw[,1]>qcrit),1,0))/n2
      sig=c(n2,beta,ICC2,ICC3,sig.raw,sig.hoch,sig.tuk, ,tot.sig,tot.sig.hoch,tot.sig.tuk)

   pow.raw=sum(ifelse(t.raw[,2]<.05,1,0))/n2
   pow.hoch=sum(ifelse(p.hoch<.05,1,0))/n2
   pow.tuk=sum(ifelse(abs(t.raw[,1]>qcrit),1,0))/n2

   pow=c(n2,beta,ICC2,ICC3,pow.raw,pow.hoch,pow.tuk)
   mat.sig[a,]=sig
   mat.pow[a,]=pow

  }

  matsig.eq[[condrep]]=mat.sig
  matpow.eq[[condrep]]=mat.pow

}

mat.sig.eq=matrix(NA,ncol=10,nrow=numcond)
mat.pow.eq=matrix(NA,ncol=7,nrow=numcond)
for(b in 1:numcond){
  means=colMeans(as.data.frame(matsig.eq[[b]]))
  power=colMeans(as.data.frame(matpow.eq[[b]]))
  mat.sig.eq[b,]=means
  mat.pow.eq[b,]=power

colnames(mat.sig.eq)=c("L2n","Effect","ICC2","ICC3","Unadjusted","Hochberg","Tuke
y","Tot Sig","Tot Hoch","Tot Tuk")

colnames(mat.pow.eq)=c("L2n","Effect","ICC2","ICC3","Unadjusted","Hochberg","Tuk
ey")
```

```
}
```

B1 Procedure

```
library("R2OpenBUGS")
for(a in 1:numbcond){
 l1n=30
 n2=l2n[conditions[a,1]]
 mat=matrix(nrow=reps,ncol=2)
 matrhat=matrix(nrow=reps,ncol=n2+2)

 for(b in 1: reps){
 dat=read.csv(paste(getwd(),"/eqvar.cond",a,".rep",b,".csv",sep=""),header=T)

 bayes1model=function() {
  for (i in 1:N) {
   y[i] ~ dnorm(y.hat[group[i]],tau.y)
  }
  for (j in 1:ngroup){
   y.hat[j] ~ dnorm(mu.a,tau.a)
  }
  mu.a ~ dnorm(0,.01)
  sigma.y ~ dunif(.0001 ,100)
  tau.y <- 1/sigma.y
  sigma.a ~ dunif(.0001 ,100)
  tau.a <-1 /sigma.a
 }

 bayes1data=list(y=dat$y,group=dat$level2,N=length(dat$y),ngroup=max(dat$level2))


bayes1out=bugs(data=bayes1data,inits=NULL,parameters.to.save=c("y.hat","mu.a","sig
ma.y"),
              model.file=bayes1model,n.chains=2,n.iter=2000,debug=F)

 bayes.means=bayes1out$summary[1:max(dat$level2),]
 sig.bm1=ifelse(bayes.means[,3]<0&bayes.means[,7]>0,0,1)
 r.hat=bayes1out$summary[1:length(bayes1out$summary[,1])-1,8]
 mat[b,]=c(ifelse(sum(sig.bm1)>=1,1,0),sum(sig.bm1/length(sig.bm1)))
 matrhat[b,]=r.hat
 }

 list.eq[[a]]=mat
 list.rhat.eq[[a]]=matrhat
 con.check.eq[[a]]=apply(matrhat,2,max)
 print(a)
```

```
}
```

B1V Procedure

```
bayes1model=function() {
    for (i in 1:N) {
      y[i] ~ dnorm(y.hat[group[i]],tau.y[group[i]])
    }
    for (j in 1:ngroup){
      y.hat[j] ~ dnorm(mu.a,tau.a)
      sigma.y[j] ~ dunif(.0001 ,UB[j])
      UB[j] <- 2*sig.samp[j]
      tau.y[j] <- 1/sigma.y[j]
    }
    mu.a ~ dnorm(0,.01)
    sigma.a ~ dunif(.0001 ,100)
    tau.a <-1 /sigma.a
  }

bayes1data=list(y=dat$y,group=dat$level2,N=length(dat$y),ngroup=length(unique(dat$l
evel2)),
                sig.samp=c(tapply(dat$y,dat$level2,sd)))

bayes1out=bugs(data=bayes1data,inits=NULL,parameters.to.save=c("y.hat","mu.a","sig
ma.y","sigma.a"),
               model.file=bayes1model,n.chains=2,n.iter=2000,debug=F)

  bayes.means=bayes1out$summary[1:length(unique(dat$level2)),]
  sig.bm1=ifelse(bayes.means[,3]<0&bayes.means[,7]>0,0,1)
  mat[b,]=c(ifelse(sum(sig.bm1)>=1,1,0),sum(sig.bm1/length(sig.bm1)))
  r.hat=bayes1out$summary[1:length(bayes1out$summary[,1])-1,8]
  matrhat[b,]=r.hat
```

B1D Procedure

```
bm0=function() {
    for (i in 1:N) {
      y[i] ~ dnorm(y.hat[i],tau.y)
      y.hat[i] <- gamma00 + delta[group[i]]
    }

    sigma.y ~ dunif(.0001 ,100)
    tau.y <- 1/sigma.y

    gamma00~dnorm(0,tau.a)
    sigma.a ~ dunif(.0001,100)
```

```
  tau.a <- 1 / sigma.a

  for (j in 1:ngroup){
   delta[j] <- 0 * equals( B[j], 1 ) + D[j] * equals( B[j], 0 )
   B[j] ~ dbern( rho[j] )
   rho[j] ~ dbeta( 1, 2 )
   D[j] ~ dnorm(0, inv_theta[j] )
   inv_theta[j] ~ dgamma( 2.1, inv_d0[j] )
   inv_d0[j] <- 1 / 0.00005
   theta[j] <- 1 / inv_theta[j]
   deltazero[j] <- equals( delta[j], 0 )
  }
 }

 bm0data=list(y=dat$y,N=length(dat$y),ngroup=max(dat$level2),group=dat$level2,
      gamma00=mean(dat$y))
 bm0.inits=function(){
  list(y.hat=rnorm(max(dat$level2)),sigma.y=runif(1),delta=rnorm(dat$level2),
      delatzero=runif(dat$level2))}

 bm0out=bugs(data=bm0data,inits=NULL,
        parameters.to.save=c("sigma.y","delta","deltazero"),
        model.file=bm0,n.chains=2,n.iter=2000,debug=F,n.thin = 3)

 bayes.means=bm0out$summary[2:(max(dat$level2)+1),]
 sig.bm1=ifelse(bayes.means[,3]<=0&bayes.means[,7]>=0,0,1)
 sig.bm2=ifelse(bm0out$summary[(max(dat$level2)+2):
      (length(bm0out$summary[,1])-1),5]==1,0,1)
 r.hat=bm0out$summary[1:length(bm0out$summary[,1])-1,8]
 mat[(b),]=c(ifelse(sum(sig.bm1)>=1,1,0),sum(sig.bm1/length(sig.bm1)),
      ifelse(sum(sig.bm2)>=1,1,0),sum(sig.bm2/length(sig.bm2)))
 matrhat[(b),]=r.hat
```

B1DV Procedure

```
bm0=function() {
   for (i in 1:N) {
    y[i] ~ dnorm(y.hat[i],tau.y[group[i]])
    y.hat[i] <- gamma00 + delta[group[i]]
   }

   gamma00~dnorm(0,tau.a)
   sigma.a ~ dunif(.0001,100)
   tau.a <- 1 / sigma.a

   for (j in 1:ngroup){
    delta[j] <- 0 * equals( B[j], 1 ) + D[j] * equals( B[j], 0 )
```

```
    B[j] ~ dbern( rho[j] )
    rho[j] ~ dbeta( 1, 2 )
    D[j] ~ dnorm(0, inv_theta[j] )
    inv_theta[j] ~ dgamma( 2.1, inv_d0[j] )
    inv_d0[j] <- 1 / 0.00005
    theta[j] <- 1 / inv_theta[j]
    deltazero[j] <- equals( delta[j], 0 )

    sigma.y[j] ~ dunif(.0001 ,UB[j])
    UB[j] <- 2*sig.samp[j]
    tau.y[j] <- 1/sigma.y[j]
  }
}

bm0data=list(y=dat$y,N=length(dat$y),ngroup=max(dat$level2),group=dat$level2,
       gamma00=mean(dat$y), sig.samp=c(tapply(dat$y,dat$level2,sd)))

bm0out=bugs(data=bm0data,inits=NULL,
           parameters.to.save=c("delta","deltazero","sigma.y"),
           model.file=bm0,n.chains=2,n.iter=2000,debug=F,n.thin = 3)

bayes.means=bm0out$summary[1:max(dat$level2),]
sig.bm1=ifelse(bayes.means[,3]<=0&bayes.means[,7]>=0,0,1)
sig.bm2=ifelse(bm0out$summary[(max(dat$level2)+1):
       ((max(dat$level2)+1)+(max(dat$level2)-1)),5]==1,0,1)
r.hat=bm0out$summary[1:length(bm0out$summary[,1])-1,8]
mat[(b),]=c(ifelse(sum(sig.bm1)>=1,1,0),sum(sig.bm1/length(sig.bm1)),
       ifelse(sum(sig.bm2)>=1,1,0), sum(sig.bm2/length(sig.bm2)))
matrhat[(b),]=r.hat
```

APPENDIX C: FULL DATA GENERATION TABLES

Table 25. *Mean Parameter Generation Results when ICC2 = .15, ICC3 = 0 for all* $\sigma^2_{ijk} = 1$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Bias	0.001	-0.007	-0.001	-0.002	-0.001	0.001	-0.002	-0.005	0.002
	SD	0.145	0.1	0.071	0.143	0.102	0.074	0.143	0.099	0.073
	RMSE	0.145	0.1	0.071	0.143	0.102	0.074	0.143	0.1	0.073
$\widehat{\beta_{ijk}}$	Mean	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	Bias	0.001	-0.007	-0.001	0.021	0.002	0.003	-0.002	-0.002	0.002
	SD	0.145	0.1	0.071	0.13	0.098	0.071	0.146	0.102	0.07
	RMSE	0.145	0.1	0.071	0.145	0.102	0.074	0.143	0.099	0.073

Table 26. *Mean Parameter Generation Results when ICC2 = .25, ICC3 = 0 for all* $\sigma^2_{ijk} = 1$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	-0.007	0.004	-0.004	0.004	-0.006	0	-0.006	0.002	-0.001
	Bias	-0.007	0.004	-0.004	0.004	-0.006	0	-0.006	0.002	-0.001
	SD	0.193	0.135	0.096	0.197	0.132	0.097	0.194	0.138	0.099
	RMSE	0.193	0.135	0.096	0.197	0.133	0.097	0.194	0.138	0.099
$\widehat{\beta_{ijk}}$	Mean	-0.007	0.004	-0.004	0.241	0.208	0.208	0.502	0.498	0.5
	Bias	-0.007	0.004	-0.004	0.041	0.008	0.008	0.002	-0.002	0
	SD	0.193	0.135	0.096	0.164	0.124	0.09	0.187	0.136	0.093
	RMSE	0.193	0.135	0.096	0.201	0.133	0.097	0.194	0.138	0.099

Table 27. *Mean Parameter Generation Results when ICC2 = .15, ICC3 = .1 for all* $\sigma^2_{ijk} = 1$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	-0.001	0.003	0.004	0.007	0.003	0.001	0.001	-0.005	-0.003
	Bias	-0.001	0.003	0.004	0.007	0.003	0.001	0.001	-0.005	-0.003
	SD	0.155	0.144	0.147	0.162	0.153	0.137	0.145	0.148	0.144
	RMSE	0.155	0.144	0.147	0.162	0.153	0.137	0.145	0.148	0.144
$\widehat{\beta_{ijk}}$	Mean	-0.001	0.003	0.004	0.213	0.209	0.204	0.507	0.5	0.508
	Bias	-0.001	0.003	0.004	0.013	0.009	0.004	0.007	0	0.008
	SD	0.155	0.144	0.147	0.132	0.118	0.122	0.14	0.132	0.133
	RMSE	0.155	0.144	0.147	0.162	0.153	0.137	0.145	0.148	0.144

Table 28. *Mean Parameter Generation Results when ICC2 = .25, ICC3 = .1 for all* $\sigma^2_{ijk} = 1$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.007	0	0.003	-0.007	0	0.002	0.007	0.007	0
	Bias	0.007	0	0.003	-0.007	0	0.002	0.007	0.007	0
	SD	0.151	0.115	0.091	0.152	0.114	0.088	0.15	0.116	0.092
	RMSE	0.151	0.115	0.091	0.152	0.114	0.088	0.15	0.117	0.092
$\widehat{\beta_{ijk}}$	Mean	0.007	0	0.003	0.235	0.215	0.212	0.499	0.493	0.502
	Bias	0.007	0	0.003	0.035	0.015	0.012	-0.001	-0.007	0.002
	SD	0.151	0.115	0.091	0.136	0.106	0.084	0.144	0.113	0.087
	RMSE	0.151	0.115	0.091	0.155	0.115	0.089	0.15	0.117	0.092

Table 29. *Mean Parameter Generation Results when ICC2=.15, ICC3=0 for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.001	-0.002	0.001	0.002	0.005	0	0.007	0.003	-0.004
	Bias	0.001	-0.002	0.001	0.002	0.005	0	0.007	0.003	-0.004
	SD	0.149	0.101	0.076	0.133	0.101	0.075	0.148	0.102	0.071
	RMSE	0.149	0.101	0.076	0.133	0.101	0.075	0.148	0.102	0.071
$\widehat{\beta_{ijk}}$	Mean	0.001	-0.002	0.001	0.215	0.208	0.2	0.5	0.498	0.501
	Bias	0.001	-0.002	0.001	0.015	0.008	0	0	-0.002	0.001
	SD	0.149	0.101	0.076	0.134	0.098	0.069	0.145	0.105	0.074
	RMSE	0.149	0.101	0.076	0.134	0.101	0.075	0.148	0.102	0.071

Table 30. *Mean Parameter Generation Results when ICC2 = .25, ICC3=0 for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.001	-0.01	0.003	-0.009	-0.004	0.001	-0.007	0	0.006
	Bias	0.001	-0.01	0.003	-0.009	-0.004	0.001	-0.007	0	0.006
	SD	0.194	0.134	0.095	0.196	0.135	0.099	0.194	0.129	0.095
	RMSE	0.194	0.134	0.095	0.196	0.136	0.099	0.194	0.129	0.095
$\widehat{\beta_{ijk}}$	Mean	0.001	-0.01	0.003	0.229	0.214	0.205	0.504	0.495	0.498
	Bias	0.001	-0.01	0.003	0.029	0.014	0.005	0.004	-0.005	-0.002
	SD	0.194	0.134	0.095	0.171	0.125	0.09	0.185	0.137	0.092
	RMSE	0.194	0.134	0.095	0.198	0.136	0.099	0.194	0.129	0.095

Table 31. *Mean Parameter Generation Results when ICC2=.15, ICC3=.1 for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	0.002	0.004	0.005	-0.003	-0.001	0.007	-0.01	-0.003	0.003
	Bias	0.002	0.004	0.005	-0.003	-0.001	0.007	-0.01	-0.003	0.003
	SD	0.152	0.148	0.149	0.15	0.151	0.145	0.149	0.15	0.143
	RMSE	0.152	0.148	0.149	0.15	0.151	0.145	0.15	0.15	0.143
$\widehat{\beta_{ijk}}$	Mean	0.002	0.004	0.005	0.221	0.21	0.205	0.499	0.498	0.499
	Bias	0.002	0.004	0.005	0.021	0.01	0.005	-0.001	-0.002	-0.001
	SD	0.152	0.148	0.149	0.13	0.121	0.124	0.137	0.134	0.132
	RMSE	0.152	0.148	0.149	0.152	0.151	0.145	0.149	0.15	0.143

Table 32. *Mean Parameter Generation Results when ICC2=.25, ICC3=.1 for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$

	β_{ijk}	0			0.2			0.5		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\gamma_{000}}$	Mean	-0.004	0.003	0.001	0.005	0.007	-0.001	0.002	-0.001	0.004
	Bias	-0.004	0.003	0.001	0.005	0.007	-0.001	0.002	-0.001	0.004
	SD	0.148	0.115	0.092	0.146	0.117	0.094	0.156	0.117	0.089
	RMSE	0.148	0.115	0.092	0.146	0.117	0.094	0.156	0.117	0.09
$\widehat{\beta_{ijk}}$	Mean	-0.004	0.003	0.001	0.245	0.231	0.21	0.513	0.505	0.501
	Bias	-0.004	0.003	0.001	0.045	0.031	0.01	0.013	0.005	0.001
	SD	0.148	0.115	0.092	0.14	0.107	0.084	0.148	0.112	0.086
	RMSE	0.148	0.115	0.092	0.153	0.121	0.095	0.156	0.117	0.089

Table 33. *Variance Generation Results for all* $\sigma^2_{ijk} = 1$ *when* β_{ijk} *= .2 and ICC3=0*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	0.997	0.997	0.999	1.002	0.997	0.999	0.999	1.001	1
	Bias	-0.003	-0.003	-0.001	0.002	-0.003	-0.001	-0.001	0.001	0
	SD	0.08	0.058	0.04	0.086	0.056	0.042	0.082	0.058	0.042
	RMSE	0.08	0.058	0.04	0.086	0.056	0.042	0.082	0.058	0.042
$\widehat{\tau^2_{u0k}}$	Mean	0.019	0.012	0.01	0.16	0.183	0.187	0.293	0.335	0.337
	Bias	0.019	0.012	0.01	-0.016	0.006	0.01	-0.04	0.001	0.003
	SD	0.023	0.013	0.009	0.109	0.074	0.051	0.186	0.122	0.088
	RMSE	0.03	0.018	0.013	0.11	0.075	0.052	0.191	0.122	0.088
$\widehat{\tau^2_{V00}}$	Mean	0.03	0.037	0.038	0.066	0.044	0.041	0.095	0.054	0.046
	Bias	0.03	0.037	0.038	0.066	0.044	0.041	0.095	0.054	0.046
	SD	0.027	0.019	0.013	0.087	0.055	0.037	0.139	0.077	0.049
	RMSE	0.04	0.041	0.04	0.109	0.071	0.056	0.168	0.094	0.067

Table 34. Variance Generation Results for all $\sigma^2_{ijk} = 1$ *when* β_{ijk} *= .5 and ICC3=0*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	1.004	0.999	0.999	1	1	0.998	1.001	0.998	0.999
	Bias	0.004	-0.001	-0.001	0	0	-0.002	0.001	-0.002	-0.001
	SD	0.083	0.06	0.042	0.084	0.059	0.042	0.084	0.058	0.041
	RMSE	0.083	0.06	0.042	0.084	0.059	0.042	0.084	0.058	0.041
$\widehat{\tau^2_{u0k}}$	Mean	0.098	0.068	0.057	0.263	0.24	0.23	0.396	0.39	0.385
	Bias	0.098	0.068	0.057	0.086	0.064	0.054	0.063	0.057	0.052
	SD	0.056	0.029	0.017	0.173	0.099	0.062	0.249	0.154	0.095
	RMSE	0.113	0.074	0.06	0.194	0.117	0.082	0.257	0.165	0.108
$\widehat{\tau^2_{V00}}$	Mean	0.2	0.232	0.242	0.214	0.233	0.248	0.246	0.24	0.241
	Bias	0.2	0.232	0.242	0.214	0.233	0.248	0.246	0.24	0.241
	SD	0.071	0.045	0.033	0.178	0.123	0.082	0.258	0.164	0.106
	RMSE	0.213	0.236	0.244	0.278	0.264	0.261	0.356	0.29	0.264

Table 35. *Variance Generation Results for all* $\sigma^2_{ijk} = 1$ *when* β_{ijk} *= .2 and ICC3=.1*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	x	x	x	0.997	1.003	1	1.001	0.999	1.002
	Bias	x	x	x	0.002	-0.003	-0.001	-0.001	0.001	0
	SD	x	x	x	0.086	0.056	0.042	0.082	0.058	0.042
	RMSE	x	x	x	0.086	0.056	0.042	0.082	0.058	0.042
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.132	0.149	0.158	0.252	0.279	0.295
	Bias	x	x	x	-0.016	0.006	0.01	-0.04	0.001	0.003
	SD	x	x	x	0.109	0.074	0.051	0.186	0.122	0.088
	RMSE	x	x	x	0.11	0.075	0.052	0.191	0.122	0.088
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.047	0.028	0.021	0.085	0.051	0.036
	Bias	x	x	x	0.066	0.044	0.041	0.095	0.054	0.046
	SD	x	x	x	0.087	0.055	0.037	0.139	0.077	0.049
	RMSE	x	x	x	0.109	0.071	0.056	0.168	0.094	0.067

Table 36. *Variance Generation Results for all* $\sigma^2_{ijk} = 1$ *when* β_{ijk} *= .5 and ICC3=.1*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{ijk}}$	Mean	x	x	x	1	0.997	1.001	1	1.001	1.001
	Bias	x	x	x	0	-0.003	0.001	0	0.001	0.001
	SD	x	x	x	0.081	0.061	0.043	0.084	0.058	0.042
	RMSE	x	x	x	0.081	0.061	0.043	0.084	0.058	0.042
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.249	0.225	0.219	0.374	0.361	0.352
	Bias	x	x	x	0.091	0.066	0.06	0.074	0.061	0.052
	SD	x	x	x	0.162	0.161	0.162	0.202	0.165	0.146
	RMSE	x	x	x	0.185	0.174	0.173	0.216	0.176	0.155
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.232	0.248	0.269	0.256	0.261	0.281
	Bias	x	x	x	0.214	0.231	0.251	0.223	0.228	0.247
	SD	x	x	x	0.21	0.206	0.196	0.185	0.142	0.109
	RMSE	x	x	x	0.3	0.309	0.319	0.29	0.269	0.27

Table 37. *Variance Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = *.2 and ICC3=0*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	0.499	0.501	0.497	0.501	0.501	0.499	0.499	0.499	0.501
	Bias	-0.001	0.001	-0.003	0.001	0.001	-0.001	-0.001	-0.001	0.001
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.001	0.001	0.003	0.001	0.001	0.001	0.001	0.001	0.001
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	1.494	1.489	1.501	1.492	1.508	1.502	1.497	1.502	1.505
	Bias	-0.006	-0.011	0.001	-0.008	0.008	0.002	-0.003	0.002	0.005
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.006	0.011	0.001	0.008	0.008	0.002	0.003	0.002	0.005
$\widehat{\tau^2_{u0k}}$	Mean	0.02	0.014	0.01	0.161	0.183	0.184	0.298	0.331	0.338
	Bias	0.02	0.014	0.01	-0.015	0.006	0.007	-0.035	-0.002	0.005
	SD	0.025	0.014	0.009	0.111	0.077	0.051	0.191	0.13	0.085
	RMSE	0.031	0.02	0.014	0.112	0.077	0.052	0.195	0.131	0.085
$\widehat{\tau^2_{V00}}$	Mean	0.031	0.036	0.039	0.066	0.047	0.04	0.09	0.056	0.044
	Bias	0.031	0.036	0.039	0.066	0.047	0.04	0.09	0.056	0.044
	SD	0.029	0.019	0.014	0.091	0.059	0.034	0.136	0.076	0.047
	RMSE	0.042	0.041	0.041	0.112	0.075	0.052	0.163	0.095	0.064

Table 38. *Variance Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = *.5 and ICC3=0*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	0.499	0.498	0.499	0.5	0.5	0.499	0.502	0.501	0.5
	Bias	-0.001	-0.002	-0.001	0	0	-0.001	0.002	0.001	0
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.001	0.002	0.001	0	0	0.001	0.002	0.001	0
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	1.499	1.504	1.498	1.5	1.506	1.499	1.495	1.504	1.499
	Bias	-0.001	0.004	-0.002	0	0.006	-0.001	-0.005	0.004	-0.001
	SD	0	0	0	0	0	0	0	0	0
	RMSE	0.001	0.004	0.002	0	0.006	0.001	0.005	0.004	0.001
$\widehat{\tau^2_{u0k}}$	Mean	0.097	0.066	0.058	0.26	0.235	0.234	0.396	0.396	0.389
	Bias	0.097	0.066	0.058	0.083	0.059	0.058	0.062	0.062	0.056
	SD	0.056	0.028	0.018	0.165	0.096	0.06	0.256	0.156	0.1
	RMSE	0.112	0.072	0.061	0.185	0.113	0.083	0.263	0.168	0.115
$\widehat{\tau^2_{V00}}$	Mean	0.201	0.238	0.242	0.214	0.239	0.244	0.24	0.233	0.239
	Bias	0.201	0.238	0.242	0.214	0.239	0.244	0.24	0.233	0.239
	SD	0.072	0.047	0.031	0.174	0.124	0.086	0.238	0.172	0.11
	RMSE	0.213	0.242	0.244	0.276	0.269	0.259	0.338	0.29	0.263

Table 39. *Variance Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = *.2 and ICC3=.1*

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	x	x	x	0.5	0.501	0.498	0.5	0.5	0.5
	Bias	x	x	x	0	0.001	-0.002	0	0	0
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0	0.001	0.002	0	0	0
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	x	x	x	1.499	1.498	1.499	1.494	1.501	1.496
	Bias	x	x	x	-0.001	-0.002	-0.001	-0.006	0.001	-0.004
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0.001	0.002	0.001	0.006	0.001	0.004
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.162	0.163	0.167	0.275	0.302	0.306
	Bias	x	x	x	0.003	0.004	0.008	-0.025	0.002	0.006
	SD	x	x	x	0.145	0.151	0.157	0.159	0.155	0.147
	RMSE	x	x	x	0.145	0.151	0.157	0.161	0.155	0.147
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.071	0.062	0.059	0.12	0.087	0.074
	Bias	x	x	x	0.053	0.044	0.042	0.087	0.053	0.041
	SD	x	x	x	0.119	0.119	0.132	0.107	0.07	0.052
	RMSE	x	x	x	0.13	0.127	0.139	0.138	0.088	0.066

Table 40. Variance Generation Results for $\sigma^2_{i1:N/2k} = .5$ and $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = .5 and ICC3=.1

	τ^2_{u0k}	0			0.176			0.333		
	N	10	20	40	10	20	40	10	20	40
$\widehat{\sigma^2_{i1:N/2k}}$	Mean	x	x	x	0.498	0.498	0.501	0.497	0.499	0.499
	Bias	x	x	x	-0.002	-0.002	0.001	-0.003	-0.001	-0.001
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0.002	0.002	0.001	0.003	0.001	0.001
$\widehat{\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk}}$	Mean	x	x	x	1.505	1.495	1.499	1.507	1.499	1.502
	Bias	x	x	x	0.005	-0.005	-0.001	0.007	-0.001	0.002
	SD	x	x	x	0	0	0	0	0	0
	RMSE	x	x	x	0.005	0.005	0.001	0.007	0.001	0.002
$\widehat{\tau^2_{u0k}}$	Mean	x	x	x	0.253	0.229	0.216	0.376	0.369	0.355
	Bias	x	x	x	0.094	0.07	0.057	0.076	0.069	0.055
	SD	x	x	x	0.161	0.165	0.16	0.203	0.169	0.151
	RMSE	x	x	x	0.187	0.179	0.17	0.217	0.182	0.161
$\widehat{\tau^2_{V00}}$	Mean	x	x	x	0.225	0.249	0.257	0.264	0.275	0.279
	Bias	x	x	x	0.207	0.231	0.239	0.231	0.241	0.246
	SD	x	x	x	0.209	0.206	0.201	0.199	0.138	0.105
	RMSE	x	x	x	0.294	0.31	0.312	0.305	0.278	0.267

Table 41. *ICC Generation Results for all* $\sigma^2_{ijk} = 1$, β_{ijk} *= .2 and* $ICC3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.046	0.047	0.046	0.178	0.182	0.184	0.267	0.274	0.273
	Bias	0.046	0.047	0.046	0.028	0.032	0.034	0.017	0.024	0.023
	SD	0.001	0.001	0	0.002	0.002	0.001	0.003	0.002	0.002
	RMSE	0.046	0.047	0.046	0.028	0.032	0.034	0.017	0.024	0.024
$\widehat{ICC3}$	Mean	0.602	0.759	0.809	0.279	0.181	0.172	0.232	0.129	0.114
	Bias	0.602	0.759	0.809	0.279	0.181	0.172	0.232	0.129	0.114
	SD	0.384	0.239	0.161	0.31	0.193	0.138	0.282	0.163	0.111
	RMSE	0.714	0.796	0.825	0.417	0.264	0.221	0.365	0.208	0.16

Table 42. *ICC Generation Results for all* $\sigma^2_{ijk} = 1$, β_{ijk} *= .5 and* $ICC3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.228	0.23	0.23	0.311	0.316	0.321	0.373	0.379	0.381
	Bias	0.228	0.23	0.23	0.161	0.166	0.171	0.123	0.129	0.131
	SD	0.001	0.001	0.001	0.003	0.002	0.001	0.003	0.002	0.002
	RMSE	0.228	0.23	0.23	0.161	0.166	0.171	0.123	0.129	0.131
$\widehat{ICC3}$	Mean	0.67	0.775	0.809	0.436	0.481	0.513	0.362	0.364	0.377
	Bias	0.67	0.775	0.809	0.436	0.481	0.513	0.362	0.364	0.377
	SD	0.172	0.085	0.053	0.288	0.182	0.11	0.303	0.201	0.125
	RMSE	0.692	0.779	0.811	0.522	0.514	0.525	0.472	0.416	0.397

Table 43. *ICC Generation Results for all* $\sigma^2_{ijk} = 1$, β_{ijk} *= .2 and* $ICC3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.178	0.183	0.18	0.264	0.274	0.271
	Bias	x	x	x	0.028	0.033	0.03	0.014	0.024	0.021
	SD	x	x	x	0.003	0.003	0.003	0.004	0.003	0.003
	RMSE	x	x	x	0.028	0.033	0.03	0.014	0.024	0.021
$\widehat{ICC3}$	Mean	x	x	x	0.314	0.251	0.232	0.256	0.195	0.179
	Bias	x	x	x	0.214	0.151	0.132	0.156	0.095	0.079
	SD	x	x	x	0.567	0.649	0.677	0.337	0.231	0.175
	RMSE	x	x	x	0.606	0.666	0.69	0.371	0.249	0.192

Table 44. *ICC Generation Results for all* $\sigma^2_{ijk} = 1$, β_{ijk} *= .5 and* $ICC3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.314	0.315	0.324	0.369	0.374	0.381
	Bias	x	x	x	0.164	0.165	0.174	0.119	0.124	0.131
	SD	x	x	x	0.003	0.003	0.003	0.003	0.003	0.002
	RMSE	x	x	x	0.164	0.165	0.174	0.119	0.124	0.131
$\widehat{ICC3}$	Mean	x	x	x	0.458	0.502	0.538	0.378	0.392	0.422
	Bias	x	x	x	0.358	0.402	0.438	0.278	0.292	0.322
	SD	x	x	x	0.33	0.326	0.306	0.301	0.221	0.169
	RMSE	x	x	x	0.487	0.517	0.534	0.41	0.366	0.364

Table 45. *ICC Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} *= .2 and* $ICC3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.048	0.048	0.047	0.18	0.182	0.181	0.267	0.273	0.273
	Bias	0.048	0.048	0.047	0.03	0.032	0.031	0.017	0.023	0.023
	SD	0.001	0.001	0	0.002	0.002	0.001	0.003	0.002	0.002
	RMSE	0.048	0.048	0.047	0.03	0.033	0.031	0.017	0.023	0.023
$\widehat{ICC3}$	Mean	0.609	0.739	0.806	0.283	0.186	0.173	0.222	0.137	0.11
	Bias	0.609	0.739	0.806	0.283	0.186	0.173	0.222	0.137	0.11
	SD	0.387	0.253	0.166	0.317	0.197	0.138	0.285	0.166	0.108
	RMSE	0.722	0.781	0.823	0.425	0.271	0.221	0.361	0.216	0.155

Table 46. *ICC Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} *= .5 and* $ICC3 = 0$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	0.228	0.232	0.231	0.311	0.316	0.321	0.371	0.377	0.382
	Bias	0.228	0.232	0.231	0.161	0.166	0.171	0.121	0.127	0.132
	SD	0.001	0.001	0.001	0.003	0.002	0.001	0.003	0.002	0.002
	RMSE	0.228	0.232	0.231	0.161	0.166	0.171	0.121	0.127	0.132
$\widehat{ICC3}$	Mean	0.674	0.783	0.806	0.438	0.491	0.503	0.363	0.352	0.372
	Bias	0.674	0.783	0.806	0.438	0.491	0.503	0.363	0.352	0.372
	SD	0.177	0.083	0.054	0.29	0.18	0.115	0.3	0.202	0.131
	RMSE	0.697	0.788	0.808	0.525	0.523	0.516	0.471	0.406	0.395

Table 47. *ICC Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = *.2 and* $ICC3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.182	0.18	0.183	0.271	0.273	0.272
	Bias	x	x	x	0.032	0.03	0.033	0.021	0.023	0.022
	SD	x	x	x	0.003	0.003	0.003	0.004	0.003	0.003
	RMSE	x	x	x	0.032	0.03	0.033	0.021	0.024	0.022
$\widehat{ICC3}$	Mean	x	x	x	0.285	0.249	0.243	0.274	0.202	0.178
	Bias	x	x	x	0.185	0.149	0.143	0.174	0.102	0.078
	SD	x	x	x	0.591	0.65	0.663	0.317	0.222	0.178
	RMSE	x	x	x	0.619	0.667	0.678	0.361	0.244	0.195

Table 48. *ICC Generation Results for* $\sigma^2_{i1:N/2k} = .5$ *and* $\sigma^2_{i\left(\frac{N}{2}\right)+1:Nk} = 1.5$, β_{ijk} = *.5 and* $ICC3 = .1$

	ICC2	0			0.15			0.25		
	N	10	20	40	10	20	40	10	20	40
$\widehat{ICC2}$	Mean	x	x	x	0.312	0.318	0.317	0.372	0.381	0.382
	Bias	x	x	x	0.162	0.168	0.167	0.122	0.131	0.132
	SD	x	x	x	0.003	0.003	0.003	0.003	0.003	0.003
	RMSE	x	x	x	0.162	0.168	0.167	0.122	0.131	0.132
$\widehat{ICC3}$	Mean	x	x	x	0.442	0.502	0.53	0.388	0.399	0.42
	Bias	x	x	x	0.342	0.402	0.43	0.288	0.299	0.32
	SD	x	x	x	0.33	0.323	0.315	0.304	0.216	0.168
	RMSE	x	x	x	0.476	0.516	0.533	0.419	0.369	0.361