A Monte Carlo Study of Several Alpha-Adjustment Procedures Used in Testing Multiple Hypotheses in Factorial Anova

A dissertation presented to
the faculty of
the College of Education of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Qian An
June 2010

© 2010 Qian An. All Rights Reserved.
This dissertation titled
A Monte Carlo Study of Several Alpha-Adjustment Procedures Used in Testing Multiple
Hypotheses in Factorial Anova

by

QIAN AN

has been approved for
the Department of Educational Studies
and the College of Education by

______________________________

Gordon P. Brooks
Associate Professor of Educational Studies

______________________________

Renée A. Middleton
Dean, College of Education
ABSTRACT

AN, QIAN, Ph.D., June 2010, Educational Research and Evaluation

A Monte Carlo Study of Several Alpha-Adjustment Procedures Used in Testing Multiple Hypotheses in Factorial Anova (197 pp.)

Director of Dissertation: Gordon P. Brooks

The Type I error rate inflates greatly when multiple hypotheses are tested using an unadjusted alpha per test procedure. Therefore, several alpha-adjustment Multiple Hypothesis Testing (MHT) procedures can be used to control the Type I error inflation while providing adequate statistical power. There are numerous statistical designs that involve MHT, such as factorial ANOVA. This study investigated the Type I error rates and statistical power rates of several alpha-adjustment MHT procedures (Bonferroni, Holm, Hochberg, and Benjamini-Hochberg (B-H)) in a balanced factorial ANOVA. Three indicators for Type I error rates were used: samplewise familywise error rate (SFWER), testwise familywise error rate (TFWER), and false discovery rate (FDR). Three criteria for statistical power rates were employed: samplewise power (SPOWER), testwise power (TPOWER), and true discovery rate (TDR). MHT procedures were also compared to the unadjusted alpha per test procedure. All statistical analyses were done with 20,000 replications as a Monte Carlo simulation in the R programming language.

Two-way and three-way fixed-effects balanced designs were analyzed. The sample size per cell was 32 in the two-way and 16 for the three-way. A medium effect size of .50 for all false null effects was used to create data with different patterns of means. MHT procedures were found to have advantages over the unadjusted alpha per
test procedure in terms of controlling the Type I error inflation at an accurate level. Specifically, Bonferroni, Holm, and Hochberg were better able to control the Type I error rates at .05. The SFWER and FDR from the B-H procedure inflate under certain conditions. The Bonferroni procedure has the lowest power while the B-H procedure has the greatest power. The Hochberg procedure worked best in this study overall. However, if the independence assumption is not met, the Holm procedure can be the best choice.

Approved: _____________________________________________________________

Gordon P. Brooks

Associate Professor of Educational Studies
ACKNOWLEDGMENTS

I would never accomplish this “tough project” (used to be impossible for me three years ago) without the great support and advice from the committee. First of all, I am forever indebted to Dr. Gordon P. Brooks, who actually saved my life. I could never forget that I felt very confused when Dr. Brooks talked about the Monte Carlo simulation for the first time. I never thought I could have done the dissertation related to this area, even using R, but finally, I made it. I really appreciate Dr. Brooks for always making time in his extremely tight schedule to provide me with amazing advice, guidance, and unforgettable support. For someone as sought after as Dr. Brooks to spare time in his day to give me encouragement and invaluable help is truly kind.

Another professor I would like to express my extreme gratitude for is Dr. George Johanson. He is patient, friendly, and provides a unique and helpful perspective on issues that allowed me to develop and expand my ideas like never before. I will never forget the encouragement that he always gave me when I was feeling a lot of pressure.

I should also show my thankfulness to Dr. Valerie Conley and Dr. Bruce Carlson. Both of you have busy schedules but still provide me timely feedback and advice. The courses that I took with them not only widened my horizon, but also taught me how to be a knowledgeable and professional person for the future. By the way, thanks Dr. Carlson for first letting me know the Holm procedure.

I also need to thank my good friend, Yanju Li. You are the nicest person that I have ever met, not only because of the delicious food have you often made for me, but also your consideration and warm-heartedness, which taught me a lot. Meanwhile, I
should show my thankfulness to Weiyu Yang and Dunke Zhou, I felt so lucky to get to know you in the fall of 2009. I could not have started the R coding without the timely help from Dunke Zhou. In fact, there are a lot of persons that I should thank, including Yonggui Zheng and Jing Ning. Thank you to all who helped me a lot during the work of my dissertation.

Finally, I need to thank my husband, my mother, and my father, who give me permanent true love and encouragement. I love you all!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>5</td>
</tr>
<tr>
<td>List of Tables</td>
<td>10</td>
</tr>
<tr>
<td>List of Figures</td>
<td>11</td>
</tr>
<tr>
<td>Chapter One: Introduction</td>
<td>12</td>
</tr>
<tr>
<td>Background of the Study</td>
<td>12</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>14</td>
</tr>
<tr>
<td>Several Kinds of Type I Error Rates</td>
<td>15</td>
</tr>
<tr>
<td>Several Kinds of Statistical Power</td>
<td>18</td>
</tr>
<tr>
<td>Factorial Analysis of Variance (ANOVA)</td>
<td>19</td>
</tr>
<tr>
<td>Monte Carlo Simulation</td>
<td>21</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>22</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>22</td>
</tr>
<tr>
<td>Delimitations and Limitations of the Study</td>
<td>25</td>
</tr>
<tr>
<td>Definitions of Terms</td>
<td>26</td>
</tr>
<tr>
<td>Organization of the Study</td>
<td>29</td>
</tr>
<tr>
<td>Chapter Two: Review of Literature</td>
<td>30</td>
</tr>
<tr>
<td>Critical Review of Relevant Philosophical Issues</td>
<td>30</td>
</tr>
<tr>
<td>Type I Error Rate</td>
<td>31</td>
</tr>
<tr>
<td>Statistical Power</td>
<td>33</td>
</tr>
<tr>
<td>Adjust Alpha or Not</td>
<td>35</td>
</tr>
<tr>
<td>Definition of the “Family”</td>
<td>40</td>
</tr>
<tr>
<td>Multiple Hypothesis Testing (MHT) Applications</td>
<td>42</td>
</tr>
<tr>
<td>Multiple Comparison in One-Way ANOVA</td>
<td>43</td>
</tr>
<tr>
<td>Multiple Tests with Correlated Variables</td>
<td>44</td>
</tr>
<tr>
<td>Multivariate Analysis of Variance (MANOVA)</td>
<td>44</td>
</tr>
<tr>
<td>Differential Item Functioning (DIF)</td>
<td>45</td>
</tr>
<tr>
<td>Multiple Regression</td>
<td>45</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Repeated Measures</td>
<td>46</td>
</tr>
<tr>
<td>Factorial Analysis of Variance (ANOVA)</td>
<td>46</td>
</tr>
<tr>
<td>Multiple Comparison Procedures in One-way ANOVA</td>
<td>47</td>
</tr>
<tr>
<td>The Tukey HSD and Tukey/Kramer Method</td>
<td>47</td>
</tr>
<tr>
<td>The Dunn-Šidák Procedure</td>
<td>48</td>
</tr>
<tr>
<td>The Scheffé Procedure</td>
<td>48</td>
</tr>
<tr>
<td>Procedures in Multiple Hypothesis Testing</td>
<td>49</td>
</tr>
<tr>
<td>Unadjusted Alpha Per Test Approach</td>
<td>49</td>
</tr>
<tr>
<td>The Bonferroni Procedure</td>
<td>50</td>
</tr>
<tr>
<td>The Modified Bonferroni Procedures</td>
<td>51</td>
</tr>
<tr>
<td>The B-H Procedure</td>
<td>53</td>
</tr>
<tr>
<td>Several Issues in Factorial ANOVA</td>
<td>55</td>
</tr>
<tr>
<td>Balanced and Unbalanced Design</td>
<td>56</td>
</tr>
<tr>
<td>Consequences of Assumption Violation in Factorial Design</td>
<td>57</td>
</tr>
<tr>
<td>Statistical Power Analysis in Factorial ANOVA</td>
<td>60</td>
</tr>
<tr>
<td>Several Related Monte Carlo Studies Mainly in Factorial ANOVA</td>
<td>61</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>64</td>
</tr>
<tr>
<td>Chapter Three: Methodology</td>
<td>68</td>
</tr>
<tr>
<td>Monte Carlo Simulation</td>
<td>69</td>
</tr>
<tr>
<td>Research Design</td>
<td>71</td>
</tr>
<tr>
<td>Monte Carlo Simulation in R</td>
<td>76</td>
</tr>
<tr>
<td>Operational Definitions</td>
<td>77</td>
</tr>
<tr>
<td>Sampling Plan and Data Generation</td>
<td>81</td>
</tr>
<tr>
<td>Data Collection Procedure</td>
<td>82</td>
</tr>
<tr>
<td>Verification of the Data Collection Procedures</td>
<td>84</td>
</tr>
<tr>
<td>Data Analysis Procedure</td>
<td>86</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>87</td>
</tr>
<tr>
<td>Chapter Four: Results</td>
<td>89</td>
</tr>
<tr>
<td>Type I Error Rate in the Two-way Balanced Design</td>
<td>89</td>
</tr>
<tr>
<td>Statistical Power Rate in the Two-way Balanced Design</td>
<td>95</td>
</tr>
<tr>
<td>Type I Error Rate in the Three-way Balanced Design</td>
<td>100</td>
</tr>
</tbody>
</table>
Statistical Power Rate in the Three-way Balanced Design ......................... 108
Summary of Type I Error Rates ........................................................................ 112
Performance of the Unadjusted Alpha Per Test Procedure .......................... 112
Performance of the Bonferroni Procedure ...................................................... 113
Performance of the Holm Procedure .............................................................. 114
Performance of the Hochberg Procedure ....................................................... 114
Performance of the B-H Procedure ............................................................... 115
Summary of Statistical Power Rates ............................................................... 115
Chapter Five: Discussion, Conclusions, and Recommendations ................. 117
Discussion of the Results .................................................................................. 117
Conclusions ....................................................................................................... 127
Recommendations ............................................................................................. 130
Future Research ............................................................................................... 131
Practical Implications ....................................................................................... 133
References ......................................................................................................... 138
Appendix A: Multiple Hypothesis Testing Procedures .................................... 152
  Unadjusted Alpha Per Test Procedure ......................................................... 153
  The Bonferroni Procedure ........................................................................... 153
  The Holm Procedure .................................................................................... 155
  The Hochberg Procedure ............................................................................. 157
  The Benjamini-Hochberg Procedure ........................................................... 159
  A Small Summary ......................................................................................... 161
Appendix B: R Code for Two-way Balanced Design ....................................... 163
Appendix C: Testing Code for the Two-way Balanced Design ......................... 193
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Possible Outcomes of m Hypothesis Tests</td>
<td>15</td>
</tr>
<tr>
<td>Table 2</td>
<td>Two-way Balanced Design</td>
<td>72</td>
</tr>
<tr>
<td>Table 3</td>
<td>Three-way Balanced Design</td>
<td>74</td>
</tr>
<tr>
<td>Table 4</td>
<td>Type I Error Rates in Two-Way Balanced Design with 20,000 Replications</td>
<td>90</td>
</tr>
<tr>
<td>Table 5</td>
<td>Statistical Power Rates in Two-Way Balanced Design with 20,000 Replications</td>
<td>96</td>
</tr>
<tr>
<td>Table 6</td>
<td>Type I Error Rates in Three-Way Balanced Design with 20,000 Replications</td>
<td>101</td>
</tr>
<tr>
<td>Table 7</td>
<td>Statistical Power Rates in Three-Way Balanced Design with 20,000 Replications</td>
<td>109</td>
</tr>
<tr>
<td>Table 8</td>
<td>Three-way ANOVA Output</td>
<td>153</td>
</tr>
<tr>
<td>Table 9</td>
<td>Using the Bonferroni Procedure</td>
<td>155</td>
</tr>
<tr>
<td>Table 10</td>
<td>Using The Holm Procedure</td>
<td>157</td>
</tr>
<tr>
<td>Table 11</td>
<td>Using The Hochberg Procedure</td>
<td>159</td>
</tr>
<tr>
<td>Table 12</td>
<td>Using The B-H Procedure</td>
<td>160</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1. Samplewise FWER in two-way balanced design with 20,000 replications. ..... 93
Figure 2. Testwise FWER in two-way balanced design with 20,000 replications. .......... 94
Figure 3. FDR in two-way balanced design with 20,000 replications. ...................... 95
Figure 4. Samplewise Power in two-way balanced design with 20,000 replications. ..... 98
Figure 5. Testwise Power in two-way balanced design with 20,000 replications. ........... 99
Figure 6. TDR in two-way balanced design with 20,000 replications. ...................... 100
Figure 7. Samplewise FWER in three-way balanced design with 20,000 replications. . 106
Figure 8. Testwise FWER in three-way balanced design with 20,000 replications. ...... 107
Figure 9. FDR in three-way balanced design with 20,000 replications. ...................... 108
Figure 10. Samplewise Power in three-way balanced design with 20,000 replications. 110
Figure 11. Testwise Power in three-way balanced design with 20,000 replications. ..... 111
Figure 12. TDR in three-way balanced design with 20,000 replications. .................... 112
CHAPTER ONE: INTRODUCTION

Background of the Study

In educational research, many statistical analyses are conducted using hypothesis tests. Often, a single hypothesis is tested. For example, an investigator might test whether a group mean differs from a specified value or if there is any significant difference between a new teaching method and a traditional one. To answer these research questions, some kinds of \( t \) tests could be adopted under a given level of significance, or \( \alpha \), which is usually .05. In statistics, if the null hypothesis is incorrectly rejected, Type I error occurs. That is, there is at most a 5% of chance that researchers can make a Type I error for a single hypothesis test. However, if researchers want to compare \( k \) new teaching methods to a traditional one, \( k \) separate hypothesis tests are often performed, each at \( \alpha = .05 \). A formula is used to indicate the probability of making at least one Type I error when multiple hypotheses are independent and true null, \( p = 1 - (1 - \alpha)^k \) where \( \alpha = .05 \) and \( k \) is the number of hypothesis tests (Hochberg & Tamhane, 1987; Maxwell & Delaney, 2000; Schochet, 2008; Stevens, 1996; Toothaker, 1993). For example, if there are four independent tests (\( k = 4 \)), the probability of finding at least one spurious impact is .19, .40 for 10 tests, and .87 for 40 tests (Schochet, 2008). Therefore, the Type I error rate is inflated. In practice, researchers often ignore the number of hypothesis tests that are conducted (Brown & Russell, 1997). As a result, researchers will make unwarranted conclusions without considering the multiple hypothesis tests that are being conducted simultaneously (Schochet, 2008).
Many researchers are familiar with post hoc multiple comparison procedures, which are able to maintain the Type I error rate at $\alpha$ when a set of comparisons is made among sample means, following a significant one-way ANOVA. These procedures, such as the Tukey method, the Tukey/Kramer (TK) method, and Scheffé’s method, can tell where significant differences exist among groups. In fact, a great deal of literature has focused on multiple comparisons and numerous post hoc procedures are available for researchers (Stevens, 1999). For example, the statistical package “SPSS 17” (SPSS Inc., 2008) currently includes 18 kinds of procedures to conducting post hoc comparisons.

“Multiple comparisons are only one instance of the use of multiple hypothesis tests in a single piece of research” (Ryan, 1959, p. 26). That is, the one-way ANOVA is only one applicable area that uses multiple hypothesis tests (MHT). In fact, MHT can be conducted in several other areas, such as multiple tests with correlated variables, multivariate analysis of variance (MANOVA), multiple chi-square tests in differential item functioning (DIF), multiple regression, repeated measures, and factorial ANOVA. Meanwhile, several kinds of MHT procedures other than post hoc procedures have been made available for researchers, such as the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure. The use of MHT procedures is to control the Type I error rate when testing multiple hypotheses simultaneously.

Several researchers have also suggested that the selection of MHT procedures is mainly based on the Type I error rate and statistical power (Hochberg & Benjamini, 1990; Kirk, 1995). As Kirk (1995) stated, “Other things equal, a researcher wants to use a procedure that both controls the Type I error rate at an acceptable level and provides
maximum power” (p. 123). “The growing awareness of the trade-off between type I and type II error rates associated with multiple comparison has had a strong effect on the philosophy and methodology of multiple comparison procedure” (Hochberg & Benjamini, 1990, p. 812). “The important issues in the area of MCPs are control of alpha and power” (Toothaker, 1993, p. 21). “Controlling the overall Type I error rate over a series of hypothesis tests is an important topic of interest to applied researchers and data analysts” (Supattathum, 1994, p. 13). As Keren and Lewis (1993) stated, “Ideally, we would like to select a method that provides a powerful test while maintaining adequate Type I error control, requires few statistical assumptions, and is easy to apply” (p. 56). Therefore, the Type I error rates and statistical power rates were investigated to evaluate several MHT procedures in this study.

Purpose of the Study

The first purpose of this study was to evaluate four MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure) in terms of Type I error rates in the balanced two-way and three-way factorial ANOVA. The second purpose was to investigate the statistical power rate of four MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure) in the balanced two-way and three-way factorial ANOVA. In addition, the Type I error rates and statistical power rates of the four MHT procedures were also compared to those in the unadjusted alpha per test procedure, in which each p value from each test is compared to .05. Finally, this study was designed to provide
researchers with empirical results about the Type I error rates and statistical power rates of MHT procedures when those procedures are applied in the balanced factorial designs.

Several Kinds of Type I Error Rates

The Type I error occurs when a true null hypothesis is falsely rejected (e.g., Stevens, 1999). There are two major kinds of Type I error rates: (a) the per test error rate (PTER), and (b) familywise error rate (FWER) (Hochberg & Tamhane, 1987; Ryan, 1959; Shaffer, 1995; Toothaker, 1993). The error rate per test is also called error rate per comparison. It is the probability that any one of the hypothesis tests is falsely rejected (Hinkle, Wiersma, & Jurs, 2003; Ryan, 1959). The FWER is the probability that at least one hypothesis is incorrectly rejected in a given family (Shaffer, 1995; Toothaker, 1993).

Another type of error rate, the False Discovery Rate (FDR), was introduced in Benjamini and Hochberg (1995). The following table is adapted from Benjamini and Hochberg (1995) which illustrated the FDR.

Table 1

<table>
<thead>
<tr>
<th>Possible Outcomes of ( m ) Hypothesis Tests</th>
<th>Accepted</th>
<th>Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Null</td>
<td>( U )</td>
<td>( V )</td>
<td>( m_0 )</td>
</tr>
<tr>
<td>Non-true Null</td>
<td>( T )</td>
<td>( S )</td>
<td>( m - m_0 )</td>
</tr>
<tr>
<td>Total</td>
<td>( m - R )</td>
<td>( R )</td>
<td>( m )</td>
</tr>
</tbody>
</table>
In Table 1, \( m_0 \) is the number of true null hypotheses, \( m - m_0 \) is the number of false null hypotheses, \( m \) is the total number of hypothesis tests. \( U \) is the number of true correct acceptances, \( V \) is the number of false rejections, \( T \) is the number of false acceptances, and \( S \) is the number of correct rejections. The total number of acceptance is \( m-R \), and \( R \) is the total number of rejections. When all hypothesis tests are true null, \( m = m_0 \).

The PTER is the expected value of \( V/m \) and usually set to .05. The FWER is the probability that the number of false rejection (\( V \)) is greater than or equal 1. FDR is the expected proportion of false rejections on true null hypotheses and the total rejections on all hypotheses. Therefore, it is expressed as:

\[
FDR = E\{\frac{V}{R} \mid R > 0\}
\]

When there are all true null hypotheses, the number of false rejections (\( V \)) equals the number of total rejections (\( R \)), the FDR is 1.0. Therefore, the FDR can’t be controlled in this case (Benjamini & Hochberg, 1995). In fact, FDR can only be used in MHT where there are some nonnull hypotheses and accordingly, the number of false rejections is different from that of total rejections.

In statistics, the Type I error rates only occur when true null hypotheses are falsely rejected. That is, the number of false rejections is only based on the true null hypotheses no matter how many false null hypotheses exist. Therefore, FWER can be thought as the expected proportion of the false rejections over the number of true null hypotheses. It can be written as:

\[
FWER = E\{\frac{V}{m_0} \mid m_0 > 0\}
\]
In a multifactor design, the number of true null hypothesis \(m_0\) will vary in different scenarios. Based on different focal units of the true null hypotheses, three specific terms are used to indicate the Type I error rate from the perspective of Monte Carlo simulation in this study. First, Samplewise FWER is the probability of detecting at least one false rejection in samples that are generated in the simulation with replications. In this case, the number of true null hypothesis is the number of samples. Second, Testwise FWER is the probability of detecting false rejections on all true null tests (there are several tests in one sample, for example, there are three hypothesis tests in one sample in the two-way design). In this case, the number of true null hypothesis is the total number true null tests. Third, FDR is the expected proportion of the number of false rejections and total rejections. Therefore, the numerator of FDR is the same as that in Testwise FWER.

For example, there are seven hypothesis tests in the three-way balanced design. Suppose there are three true null hypotheses and four false null hypotheses. In the perspective of Monte Carlo simulation, a sample is generated in each replication. To calculate Samplewise FWER, the number of at least one false rejection which is only based on true null hypotheses is used to divide the total number of samples. If there are 20,000 replications in the simulation, there are 20,000 samples. Suppose 500 samples have false rejections, the Samplewise FWER is 500/20,000, or .025. If there are 20,000 iterations, there will be 60,000 true null hypothesis tests and 80,000 false null hypotheses (140,000 total hypothesis tests) in the simulation because each sample has three true null hypotheses and four false null. Suppose the false rejections are 1,000 within all true null
tests, therefore, the Testwise FWER is 1,000/60,000, or .017. Suppose there are total 30,000 rejections, the FDR is 1,000/30,000, or .03.

Several Kinds of Statistical Power

“The term power refers to the probability of rejecting a false null hypothesis” (Kirk, 1995, p. 58). It is the ability to detect a significant effect when a hypothesis test is false. There are several kinds of power that were introduced by researchers in the context of multiple comparison (e.g., Kirk, 1995; Kromrey & Dickinson, 1995; Toothaker, 1993): (a) any-pairs power is the probability of finding at least one true significant result in a set of hypotheses; (b) all-pairs power is the probability of finding all significant results in a set of hypotheses; and (c) per-pair power is the probability of finding the significant result based on each false null hypothesis in a set of hypotheses. In addition, as Kromrey and Dickinson (1995) suggested, the average of per-pair power that is based on all of the false null hypotheses in a set of hypotheses may be used as an index to indicate an average per pair power. It is necessary to mention that the terminology of “pairs” is for multiple comparison procedure (MCP) and some slightly more general terms will be used in this study, because in the context of factorial ANOVA, the power is obtained based on all hypothesis tests of the main effects and interaction effects.

As indicating the Type I error rates in this study, three specific terms in the perspective Monte Carlo simulation are used to evaluate the statistical power of MHT procedures in this study. First, Samplewise Power is the probability of detecting at least one significant effect in samples that are replicated in the simulation. Samplewise Power can be regarded as any-pair power in terms of samples in simulation. Second, Testwise
Power is the probability of detecting significant results within all false null hypotheses tests that are replicated in the simulation. Testwise Power is thought as the average of per-pair power based on all false null tests, which was suggested in Kromrey and Dickinson (1995). Third, True Discovery Rate (TDR) is the expected proportion of the number of true rejections on false null hypothesis tests and the total number of rejections in all hypothesis tests. Therefore, the numerator in the TDR is the same as that in the Testwise Power. The denominator in the TDR is the same as that in FDR. According to Table 1, the TDR can be expressed as the following:

$$TDR = E\left\{ \frac{S}{R} \mid R > 0 \right\}$$

It is necessary to note that the FDR and TDR are complementary to each other. That is, the sum of FDR and TDR is 1.0 (because they have the same denominator ($R$) and $V+S=R$). Using the same example above, there are seven tests in the three-way balanced design, in which three are true null hypotheses and four false nulls. The power calculation is actually only based on false null hypotheses. Suppose 20,000 iterations and therefore, 20,000 samples are generated and 80,000 (20,000*4) false null tests are in samples. If there are 16,000 samples with rejections, the Samplewise Power is 16,000/20,000 or .80. If there are 29,000 true rejections, the Testwise Power is 29,000/80,000, or .36. However, there are 30,000 total rejections; the TDR is 29,000/30,000, or .97.

**Factorial Analysis of Variance (ANOVA)**

“A factorial design is one in which all possible combinations of the levels of two or more treatments occur together in the design” (Kirk, 1995, p. 364). It is used to test the
effect of two or more independent variables (e.g., gender, treatments, and groups) on a set of dependent variables (Stevens, 1996). For example, if there are two independent categorical variables and each has two levels, the analysis is called a 2x2 design. Factorial designs have been commonly used in the educational and psychological areas (Vallejo, Ato, Fernandez, & Livacic-Rojas, 2008). There are some advantages in conducting factorial ANOVA over using several one-way ANOVAs at one time (Stevens, 1999). First, it enables researchers to investigate the interaction effect because there is more than one independent variable in the design, and the independent variables may interact with each other. Second, factorial designs are more powerful than one-way ANOVA. Third, it only requires as fewer as many participants in the two-way analysis as that in two one-way analyses.

Many researchers have been realizing the inflation of Type I error rate in one-way ANOVA. However, minimal attention has been received for the multi-way ANOVA or factorial ANOVA by researchers (Fletcher, Daw, & Young, 1989; Halderson & Glasnapp, 1974; Kromrey & Dickinson, 1995; Smith, Levine, Lachlan, & Fediuk; 2002). For example, in a three-factor fixed-effects design, there are total seven hypotheses tests. Based on the formula of \( p = 1 - (1 - \alpha)^k \), using alpha .05 and assuming all the hypotheses are true null and independent, the researcher can make .30 familywise Type I error. In a five-factor fixed-effects design, there are 31 hypotheses tests. If using alpha .05 and assuming all the null hypotheses are true and independent, the researcher can make .80 familywise Type I error. Obviously, the Type I error rate inflates greatly in
factorial ANOVA design. Several MHT procedures can be used to control the Type I error inflation in factorial ANOVA.

In factorial designs, if each cell has the same number of participants, it is called a balanced or orthogonal design (Stevens, 1999). It should be noted that each effect is uncorrelated in the balanced design, which is the unique characteristic in factorial design compared to other areas which also involve MHT. For example, the hypothesis tests in a correlation matrix are correlated. In this study, some MHT procedures have the assumption of independence; therefore, the balanced factorial designs are suitable and the generalizability of this study can be guaranteed accordingly.

Monte Carlo Simulation

The Monte Carlo method is based on repeated random sampling to compute results (Brooks, 1998). Mooney (1997) stated that “Monte Carlo simulation offers an alternative to analytical mathematics for understanding a statistic’s sampling distribution and evaluating its behavior in random samples” (p. 2). Monte Carlo simulation can be used to evaluate the quality of statistical procedures, such as the Type I error rate and statistical power (Mooney, 1997). For example, the researcher can draw sample data with many replications in simulation, some where the null hypotheses are true and some are false. Rejections could be counted with many iterations and the Type I error rate under the true null hypotheses would be obtained. Meanwhile, statistical power rate can be obtained with many replications under false null hypotheses. Then, researchers could compare the Type I error rates and statistical power rates of these statistical procedures.
Statement of the Problem

This study investigated the unadjusted alpha per test procedure and several alpha-adjustment MHT procedures: the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure in the context of factorial ANOVA (see Appendix for formal description and example of the procedures). Three criteria (Samplewise FWER, Testwise FWER, and FDR) were used to indicate the Type I error rates. Also three indicators (Samplewise Power, Testwise Power, and TDR) were used to show statistical power rates. By conducting a Monte Carlo simulation, the Type I error rates and statistical power rates of these procedures were obtained. More specifically, there were two primary research questions:

Research Question 1 (RQ1): How well do different MHT procedures control Type I error in terms of Samplewise FWER, Testwise FWER, and FDR in (a) the balanced two-way factorial ANOVA and (b) the balanced three-way factorial ANOVA?

Research Question 2 (RQ2): How well do different MHT procedures show statistical power rate in terms of Samplewise Power, Testwise Power, and TDR in (a) the balanced two-way factorial ANOVA and (b) the balanced three-way factorial ANOVA?

Significance of the Study

Being an important statistical design, factorial ANOVA is broadly used in educational research. Most important, the uncorrelated effects in the balanced factorial designs allow this study to include some MHT procedures which require the assumption of independence, therefore, this study can have more generalizability.
In previous literature, researchers have made great contributions to MCPs to control the Type I error inflation in one-way ANOVA, however, the inflation of Type I error rates in factorial ANOVA is often not recognized in behavioral science (Halderson & Glasnapp, 1974; Kromrey & Dickinson, 1995). Even though Keppel (1991) noted that it is common practice in psychology to disregard the increase in familywise error rate associated with the tests in factorial ANOVA, more discussions and considerations are needed. It is necessary for researchers to consider the Type I error rate inflation in factorial ANOVA (Smith, Levine, Lachlan, and Fediuk, 2002).

Three main extensions based on previous studies were presented in this empirical study. The first extension is that the Type I error rates and statistical power rates were obtained from several aspects. That is, specific criteria were used to indicate the Type I error rate and statistical power rate from the perspective of Monte Carlo simulation. For example, the Samplewise FWER, the Testwise FWER, and FDR were used to evaluate the Type I error rates of MHT procedures. The Samplewise Power, the Testwise Power, and TDR were criteria to indicate the statistical power rate. The samplewise criteria (Samplewise FWER and Samplewise Power) focus on the FWER and statistical power in the context of samples that were generated with many replications in the simulation. The testwise criteria (Testwise FWER and Testwise Power) are from the perspective of all tests in the factorial design. The FDR and TDR are complementary to each other.

The second extension is that a relatively newer MHT approach, the B-H method which is based on the false discovery rate (FDR), was included. Being an approach created in a new perspective and easily applied, the B-H method has been paid more and
more attention by researchers and become a standard method in several handbooks of some organizations, such as What Works Clearinghouse (2008) and National Center for Education Statistics (NCES) (2009). In particular, NCES suggested using the Bonferroni method in 1996 when testing multiple hypotheses, and then also suggested the B-H approach in 1998. However, not many studies have compared this more recent approach to some older ones in situations other than multiple comparison procedures and none in factorial ANOVA. In this sense, it is very necessary to evaluate such procedure in a Monte Carlo study, and the current study could fill some gaps.

Third, this study makes some connections between FWER and FDR from the practice point of view. FWER and FDR were mainly viewed from different philosophical perspective in literature (Benjamini & Hochberg, 1995; Schochet, 2008), “any procedure controls the FWER also controls the FDR” (Benjamini & Hochberg, 1995, p. 291). In addition, the control of FDR indicates controlling FWER in a weak sense when all null hypotheses are true (Benjamini & Hochberg, 1995). Many researchers are more familiar with FWER than FDR, therefore, if controlling FWER, what impact it will have on FDR. Or if controlling FDR, what impact it will have on FWER. In fact, few studies investigated these two concepts together in the context of factorial ANOVA with one Monte Carlo study. In this sense, this study can fill some gaps in terms of building some connections between FWER and FDR, which enable researchers to have a general idea about the impact on FWER and FDR when using some alpha-adjustment MHT procedures in practice.
Delimitations and Limitations of the Study

Admittedly, there were some boundaries in this study. First, this empirical study only focused on the fixed-effects factorial ANOVA. The designs were limited to the balanced two-way and three-way. In fact, Stevens (1999) pointed out that the four-way or more is not common or even more complicated and interpreting higher order interactions is very difficult. It is very rare to encounter factorial designs with more than three independent variables in related literature and there is too much complexity behind this kind of design (Keppel, 1991). Therefore, the results in this study won’t apply to the unbalanced design. Meanwhile, the Hochberg procedure and the B-H procedure have an independence assumption, which wouldn’t be appropriate in the unbalanced design.

Second, a number of MHT procedures are currently available for researchers; however, only four MHT procedures were compared to the unadjusted method in this study. The unadjusted alpha approach is most familiar to researchers, which requires each $p$ value to be compared to the nominal alpha level, usually .05, to test the significance. The Bonferroni method has become a general and standard approach in controlling the Type I error rate without requiring any constraining assumptions (Hochberg & Benjamini, 1990; Olejnik, Li, Supattathum, & Huberty, 1997). The Holm procedure and the Hochberg procedure are more powerful than the Bonferroni procedure and they both are logical “sequentially rejective” Bonferroni procedures (Hochberg, 1988; Holm, 1979). These two methods are early approaches in conducting the Bonferroni procedure sequentially (Hancock & Klockars, 1996). Although several modified B-H approaches
have been made available for researchers since 1995, only the original B-H method was included in this study, not any extensions and adaptations.

Third, this study used Monte Carlo methods, which require the researcher to make certain decisions that might limit the generalizability. First, the patterns of true null and false null effects that were used in the simulation are some subsets of all possible patterns of true null and false null effects, which seem arbitrary. But since all hypotheses are regarded as one family in this study, it doesn’t matter whether which main effect or interaction effect is significant. Some researchers suggested that the number of true null or false null effects should be emphasized rather than the specification of true null and false null effect, as far as the normal theory is concerned (Sawilowsky, Blair, & Higgins, 1989; Shaffer, 1977). Therefore, the patterns can guarantee the generalizability of current study. Third, medium effect size of .50 was mainly used for all patterns and the sample size per cell is fixed (there are 32 participants per cell in the balanced two-way analysis and 16 for the three-way), which ensure the power of .80. Fourth, the simulation was conducted through drawing samples from infinite numbers of populations with replacement instead of finite numbers without replacement. Fifth, all statistical analyses were conducted assuming the assumptions in factorial ANOVA are not violated.

Definitions of Terms

*Type I Error Rate*

“Type I error rate is the probability of rejecting the null hypothesis when it is true, or saying the group differ when they don’t” (Stevens, 1999, p. 9). Type I error is usually set to .05 or .01.
**Familywise Error Rate (FWER)**

The FWER is the probability that at least one hypothesis is incorrectly rejected in a given family (Shaffer, 1995; Toothaker, 1993).

**Experimentwise Error Rate (EWER)**

The EWER is the probability of making at least one Type I errors in a set of tests in one experiment (Kline, 2004).

**Samplewise Family Error Rate (SFWER)**

The SFWER is the probability that at least one hypothesis is incorrectly rejected in the sample that are replicated in a Monte Carlo simulation.

**Testwise Family Error Rate (TFWER)**

The TFWER is the probability of making false rejections on all true null hypothesis tests in the samples that are replicated in a Monte Carlo simulation.

**False Discovery Rate (FDR)**

FDR is the expected proportion of the number of false rejections and the total rejections (Benjamini & Hochberg, 1995).

**Type II Error**

“Type II error is the probability of accepting $H_0$ when it is false, i.e., saying the group don’t differ when they do” (Stevens, 1999, p.122).

**Statistical Power**

“It is the probability of rejecting $H_0$ when it is false. Thus, power is the probability of making a correct decision” (Stevens, 1999, p. 122). It is 1 minus Type II error in statistics. In addition, the alpha level and power are related.
Samplewise Power (SPOWER)

The SPOWER is the probability of detecting at least one significant effect based on false null hypotheses in samples that are replicated in a Monte Carlo simulation.

Testwise Power (TPOWER)

The TPOWER is the probability of detecting the significant results based on all false null hypothesis tests that are replicated in a Monte Carlo simulation.

True Discovery Rate (TDR)

The TDR is the expected proportion of the number of true rejections on false null hypothesis tests and the total rejections. The FDR and TDR are complementary to each other.

P values and Adjusted p values

The p values are supportive evidence indicating the significance of statistical estimates. “A p-value is the probability of obtaining a value of the test statistic that is equal to or more extreme than the one observed. Usually p values are obtained with the aid of a computer statistical package” (Kirk, 1995, p. 65). The p values are adjusted in MHT procedures to control the Type I error rate in a given family and therefore, they are called as adjusted p values.

Multiple Hypothesis Testing and MHT procedures

When two or more hypotheses are tested simultaneously, we can say multiple hypothesis testing (MHT). Usually, some procedures based on alpha-adjustment are used to control the Type I error inflation and we can call these procedures as MHT procedures.
Organization of the Study

There are five main Chapters in this study. Chapter one introduces the background, the purpose, the research questions, the significance, and delimitation and limitations of this study. Chapter two reviews several philosophical issues about adjusting alpha in MHT. Some MHT procedures are illustrated and several Monte Carlo studies that are related to this study are also reviewed. Chapter three mainly focuses on research design, Monte Carlo simulation in R program, data sampling, data collection, and data analysis. Chapter four reports the main results from Monte Carlo simulation. Discussion, conclusions, and recommendations are in Chapter five. References and Appendices (e.g., MHT procedures and R code) are attached at the end of this study.
CHAPTER TWO: REVIEW OF LITERATURE

Critical Review of Relevant Philosophical Issues

The rationale behind multiple hypothesis testing (MHT) procedures is to adjust the alpha level so as to control the Type I error rate while providing adequate power. As several researchers suggested, a MHT procedure not only can control the Type I error rate at an acceptable level, but also can provide adequate statistical power (Hochberg & Benjamini, 1990; Hancock & Klockars, 1996; Kirk, 1995). “The growing awareness of the trade-off between type I and type II error rates associated with multiple comparison has had a strong effect on the philosophy and methodology of multiple comparison procedure” (Hochberg & Benjamini, 1990, p. 812). The important issues in MTH are the Type I error control and statistical power (Toothaker, 1993). Therefore, both the Type I error rates and statistical power are investigated for several MHT procedures in this study.

“One goal of any researcher is to minimize both Type I and Type II errors” (Hinkle, Wiersma, & Jurs, 2003, p. 300). The concept of Type I error rate, Type II error rate and statistical power will be reviewed at the beginning of this Chapter. In the literature, there is a lack of consensus about whether or not to adjust the alpha level in social behavior science. A review of reasons of adjusting alpha or not will be offered. The purpose of this study is to investigate the Type I error rate and statistical power of four MHT procedures in a fixed-effect factorial ANOVA design. A review of MHT procedures will also be included. The procedures primarily discussed are the classical Bonferroni procedure, modified Bonferroni procedures, and the B-H approach. Even
though much of the literature is concerned with multiple comparison procedures (MCP), this literature also applies generally to MHT procedures. As Ryan (1959) stated, “Multiple comparisons are only one instance of the use of multiple hypothesis tests in a single piece of research” (p. 26). In addition, some specific issues about the design of factorial ANOVA, such as orthogonal and non-orthogonal design, consequences of violating assumptions, statistical power in factorial ANOVA, and definition of family will be reviewed. Several Monte Carlo studies that are related to the current study will also be reviewed. A summary will be included at the end of this chapter.

Type I Error Rate

Type I error occurs when a true null hypothesis is falsely rejected. It is equal to the significance level, \( \alpha \) (e.g., Cohen, 1973, 1988; Field, 2005; Stevens, 1999). The Type I error rate, or alpha (\( \alpha \)) is regarded as a kind of significance criterion (Cohen, 1973). As he stated, “The decision- or significance-criteria, \( \alpha \): This is the expression of the researcher’s policy with regard to risking the mistaken rejection of \( H_0 \) in the form of a long-term error rate for rejecting when \( H_0 \) is true” (Cohen, 1973, p. 225).

There are two major kinds of Type I error rates: (a) the per test error rate (PTER), and (b) familywise error rate (FWER) (Hochberg & Tamhane, 1987; Ryan, 1959; Shaffer, 1995; Toothaker, 1993). The error rate per test is also called error rate per comparison. It is the probability that any one of the hypothesis tests will be falsely rejected (Hinkle, Wiersma, & Jurs, 2003; Ryan, 1959). The FWER is the probability that at least one hypothesis is incorrectly rejected in a given family (Shaffer, 1995; Toothaker, 1993).
Ryan (1959) stated that “It is of much greater practical importance to consider which of the error rates is the best representation of the dependability or ‘significance’ of our results” (p. 35). There has been a lot of literature that focused on the question of what kind of error rate should be chosen. Ryan (1959) suggested that careful analysis be required to decide the error rate. Miller (1981) generally recommended FWER control in MCP. Controlling FWER is appropriate in terms of holding the overall Type I error probabilities in post hoc procedures (Hochberg & Tamhane, 1987; Keppel, 1991).

Some researchers justified several kinds of error rates: Error rate per test (PTER) and familywise error rate (FWER) (Hancock & Klockars, 1996; Maxwell & Delaney, 2000; Toothaker, 1993). These researchers showed that PTER has large power at the expense of high Type I error. FWER has some advantages over PTER. Controlling FWER is actually also controlling PTER. That is, if FWER is kept small, then PTER is also kept small. Hancock and Klockars (1996) argued that PTER offers the most power to detect a false null hypothesis, but this approach could inflate the Type I errors within a set of comparisons. Therefore, a more conservative concept unit, the FWER, was suggested because the familywise alpha ($\alpha_{FW}$) does not exceed the nominal level. Keppel (1991) contended that the familywise Type I error becomes relevant when more than one statistical test is conducted in the analysis of an experiment. He stated that “familywise error is the inevitable penalty associated with conducting additional comparisons” (p. 247). Hancock and Klockars (1996) also stated that “The very existence of the plethora of MHT research implies that a number of researchers consider a familywise approach to be more appropriate” (p. 272). PTER and FWER would be the same when there is only
one test in an experiment. However, when there is more than one test, the decision should be based on the nature of the hypothesis tests (Kirk, 1995).

*Statistical Power*

Type II error is the probability of accepting the null hypotheses when they are actually false (Aitza, 1993; Field, 2005; Hinkle, Wiersma, & Jurs, 2003; Stevens, 1999). The Type I error and Type II error are regarded as two main kinds of inference error (Aitza, 1993). In practice, the Type II error is difficult to calculate (Aitza, 1993; Field, 2005; Hinkle, Wiersma, & Jurs, 2003; Stevens, 1999). The complement of the Type II error is power. As Cohen (1988) stated, “The power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists” (p. 4). The Type I error, or alpha, Type II error, and statistical power are closely related (Aitza, 1993; Cohen, 1988). They are significance criterion in hypothesis tests (Cohen, 1988; Rossi, 1990).

The importance of reporting explicit power has been argued by researchers since 1970s (e.g., Brewer, 1972; Cohen, 1973). “It is almost universally accepted by educational researchers that the power of a statistical test is important and should be substantial” (Brewer, 1972, p. 391). As Rossi (1990) stated, “Because the aim of behavioral research is to discover important relations of variables, a consideration of power might be regarded as a natural and important part of the planning and interpretation of the research” (p. 646). There are several reasons that were suggested by researchers of reporting statistical power (Rossi, 1990): First, it enables researchers to
know the probability of obtaining a statistically significant result. Second, it facilitates the explanation of null results. That is, the failure of rejecting a null hypothesis doesn’t mean the hypothesis is really true, it is because under lower power. Third, “statistical power provides insight concerning entire research domains” (Rossi, 1990, p. 646). A statistically significant result may be questioned in a design with a low power. As Kline (2004) stated, “the concept of power does not stand by itself without hypothesis tests” (p. 43).

Cohen (1988) made great contribution in terms of power analysis for different statistical designs. Researchers are able to determine the statistical power for a design based on this book. For example, given the alpha level and the magnitude of effect size, appropriate sample size that are required in factorial ANOVA can ensure enough power, for example, .80. There are several factors that influence the statistical power in a statistical analysis: the significance level alpha ($\alpha$), the effect size, and the sample size. Usually, the alpha is set to .05 or .10. The effect size is mainly based on the Cohen’s (1988) in which different effect size for different designs were illustrated and categorized. Appropriate sample size in statistical designs can ensure enough power, and the power of .80 is commonly used which can give researchers enough confidence to make conclusions from significant results. If holding other factors constant, the less alpha level, the less powerful test; the larger sample size, the more powerful test; the larger effect size, the more powerful test (Cohen, 1988; Hinkle, Wiersma, & Jurs, 2003; Levin, 1997; Tabachnick & Fidell, 2001; Stevens, 1999).

There are several kinds of power that were introduced by researchers in the context of MCP (e.g., Kirk, 1995; Kromrey & Dickinson, 1995): (a) any-pairs power is
the probability of finding at least one true significant result in a set of hypotheses; (b) all-pairs power is the probability of finding all significant results in a set of hypotheses; and (c) per-pair power is the probability of finding the significant result based on each false null hypothesis in a set of hypotheses. In the context of multiple comparison, the power is the probability of rejecting a false null hypothesis (Kirk, 1995). In fact, “the power indexes used in consideration of multiple comparisons of means have direct analogies in consideration of any set of hypothesis tests” (Kromrey & Dickinson, 1995, p.54).

As there are different opinions about the concept unit for Type I error rate, the similar problem happens to the power: which kind of power is appropriate. As Kirk (1995) stated, “Consequently, when researchers investigate the relative power of multiple comparison procedures, it is customary to report data for each kind of power” (p. 124). Several researchers included the three kinds of power of MHT procedures in their Monte Carlo studies (Kromrey & Dickinson, 1995; Olejnik, Li, Supattathum, & Huberty, 1997; Supattathum, 1994).

Adjust Alpha or Not

Multiple Hypothesis Testing (MHT) is one of the most confusing statistical areas and is receiving many different arguments from researchers (Games, 1971). Hochberg and Benjamini (1990) regarded the multiplicity problem of MHT as “an incompatibility between some formal statistical methodology and existing research practice” (p. 1). There are some differing opinions in terms of whether or not to adjust alpha. Some researchers have the attitude of absolutely not adjusting alpha. As O’Keefe (2003) stated, “adjusting the alpha level because of the number of tests conducted in a given study has no
principled basis, commits one to absurd beliefs and policies, and reduces the statistical power, the practice of requiring or employing such adjustments should be abandoned” (p. 444). That is, the unadjusted alpha per test procedure should be always used. There are some key reasons of not adjusting alpha were summarized in Hancock and Klockars (1996). For example, although the unadjusted approach leads to more Type I errors, the risk is worthwhile because science becomes meaningless without any risk: the knowledge of researchers is accumulated through the risk and uncertainty. The unadjusted approach leads to consistent results across studies (O’Keefe, 2003; Wilson, 1962). The unadjusted procedure has more power to detect significant difference. As some researchers argued, the use of MHT procedures to control the Type I error rate lead to the loss of power and consequently, important treatment effects may not detected (Rothman, 1990; Saville, 1990; Wilson, 1962).

Some researchers responded to O’Keefe (2003) and argued that the alpha should be adjusted in the appropriate contexts. Hewes (2003) argued that the alpha adjustment itself is not absurd but the inappropriate use of alpha-adjustment procedures. For example, it doesn’t make any sense to calculate the mean of male and female in research. That is, it is irrational for O’Keefe (2003) to propose the absurdity of the adjustment principle. As Hewes (2003) stated, “the moral of this story is that one cannot make internally logical rules like addition, averaging, or even experiment-wise error correction absurd because we can use them for irrational goals or in inappropriate contexts” (p. 449). Alpha-adjustment procedures are statistical tools instead of a recipe (Hewes, 2003). Tutzauer (2003) contended that there are certainly some occasions that require the control
of familywise Type I error, that is, the alpha should be adjusted. For example, “when examining an effect in many subpopulations or when there is a web of claims made by a strong theory” (p. 456). Also, in a hierarchical model, where the invalidation of any step in the hierarchy influences the entire context, or “in a critical test of competing theories in which one wants to make a very strong case that the surviving theory’s claims are all true, and the nulls are not” (p. 456).

Several researchers focused on the adjustment or not in terms of planned and unplanned comparisons. For example, the planned comparisons do not require alpha adjustment but the unplanned comparisons do (Keppel & Zedeck, 1989). Cook and Farewell (1996) contended that multiplicity adjustments may not be necessary if researchers have a prior interest in estimating separate treatment effects in a given study with a limited number of tests. Hancock and Klockars (1996) illustrated that the nonorthogonal contrasts have more Type I error than the orthogonal contrasts. “The degree of redundancy within a set of contrasts has potentially powerful implications for choosing the critical values by which those contrasts are tested” (p. 278). That is, the extent of alpha-adjustment is different in the orthogonal and nonorthogonal contrasts. “a more stringent critical value needs to be chosen to invoke familywise control when each orthogonal test is tested” (p. 278). However, if the orthogonal contrasts are completely redundant (i.e. identical), the unadjusted procedure should be used because the contrasts are actually identical. As for nonorthogonal contrasts, different alpha-adjustment procedures can be used based on the extent of redundancy (Hancock & Klockars, 1996).
Many researchers are supportive of adjusting alpha in research. As Hancock and Klockars (1996) stated, “the very existence of the plethora of MCP research implies that a number of researchers consider a familywise approach to be more appropriate” (p. 272). For example, researchers argued that treatment differences are often exaggerated without adjusting alpha in MHT (Benjamini & Hochberg, 1995; Ludbrook, 1998; Schochet, 2008). Some researchers even claimed that the Type I error rate inflation must be guarded and adjusted at all cost because of leading to misunderstanding of study findings (Games, 1971; Keselman, Cribbie, & Holland, 2002; Ryan, 1959, 1960; Westfall & Young, 1993). Some philosophical issues about MHT were proposed in Tukey (1991) and he argued that the philosophy of MHT is to control the error rate. Statistical procedures are useful tools which should be used in research to control the error rate, but the solution to Type I error inflation is not unique; it depends on how researchers analyze and access the results. Both philosophical and statistical reasons to implement MHT were provided in Hancock and Klockars (1996). In particular, they stated that “The philosophical component involves the articulation of a researcher’s position on the balance between statistical power and control over type I error” (p. 270). The statistical aspects of MHT are mainly employing statistical procedures to control the Type I error rate. As they expressed that, “the quest to maximize experimental power while control over the familywise error rate is maintained” (Hancock & Klockars, 1996, p. 270).

Ryan (1959) preferred to adjust alpha in MHT and discussed some reasons why researchers neglect adjusting alpha in the context of MHT. First, there are different procedures under different assumptions, and “there are important questions of logic
involved in the use of these methods and these issues have not been clearly faced in the psychological literature” (p. 26). Second, Tukey (1953) was first to propose adjusting alpha in MHT, however, his work was not widely circulated. Third, the opinion of adjusting alpha was provided by statisticians and researchers in other areas may seldom access to these works.

In practice, it is currently very common for researchers to collect huge datasets and make many hypotheses simultaneously within one dataset. Researchers care about individual significance among a family of hypotheses in most applications (Benjamini & Yekutieli, 2001). That is, researchers need to make correct inferences based on significant results among a set of hypotheses. In this sense, controlling the Type I error using some MHT procedures is a must (Games, 1971; Hancock & Klockars, 1996; Seaman, Levin, & Serlin, 1991). Some useful guidelines have been made available for researchers. For example, Schochet (2008) listed several guidelines for multiple testing in impact evaluations of educational interventions, such as, “the multiple comparisons problem should not be ignored, limiting the number of outcomes and subgroups forces a sharp focus and is one of the best ways to address the multiple comparisons problem, the multiple comparisons problem should be addressed by first structuring the data” (p. 3). The handbook of “What Works Clearinghouse” (2008), which aims to provide the trusted and scientific evidence for the Institute of Education Science, recommended an adjustment when multiple outcomes occur or multiple comparisons are required. The NCES (2009) also suggested researchers adjust alpha and control the Type I error rate by choosing the Tukey method, the Scheffé method, and the B-H method based on research
purposes and designs. The Institute of Education Sciences at the U.S. Department of Education also developed guidelines to deal with multiple testing in education research. The B-H method is recommended in the study of some researchers because it is less conservative than other FWER methods in terms of power (Schochet, 2008; Williams, Jones, & Tukey, 1999).

**Definition of the “Family”**

Whether adjusting alpha or not involves philosophical issues and “a position of compromise may exist in the ambiguity of the term family” (Hancock & Klockars, 1996, p. 274). That is, the definition of the family affects how researchers adjust alpha. What constitutes a family has become a controversial question for researchers in the past several decades. It is particularly difficult to define a family (Hancock & Klockars, 1996). Tukey (1953) first introduced the term of family. Ryan (1959) contended that the entire experiment should constitute the family. In addition, the family is equivalent to the experiment in the one-dimensional case. However, these two terms are different if in a two-variable analysis. There would be two families in the experiment. Therefore, the familywise error rate would be different from experimentwise error rate in this sense. Miller (1981) argued that the family of hypotheses is all those actually tested on the results from a single experiment. Hochberg and Tamhane (1987) defined that “Any collection of inferences for which it is meaningful to take into account some combined measure of errors is called a family” (p. 5). They argued that the content of a family depends on the type of study. Shaffer (1995) reviewed prior literature about the definition of family and suggested different family configurations should be based on different
purposes of studies. Kirk (1995) stated that “a family of contrasts consists of those contrasts that are related in terms of their content and intended use” (p. 120). Ludbrook (1998) provided a more specific definition, a family should include “all those experimental observations that were, or could have been, analyzed statistically by global procedures” (p. 1033).

Hancock and Klockars (1996) concluded that the definition of the family seems more agreeable in the context of one-way ANOVA. However, when it comes to multifactor designs, it is less consistent. In their opinions, the definition should be expanded because a single experiment could consist of multiple tests on multiple variables in factorial designs. That is, only defining all tests in one family is not that clear. The key point is how to define a family of hypothesis tests. Kirk (1995) treated different significant effects as separate families. “The usual practice is to test treatments A and B and the A x B interaction each at the $\alpha$ level of significance” (p. 123). Maxwell and Delaney (2000) explicitly proposed that each main effect and interaction should constitute their own family. However, several researchers also suggested treating all tests in the factorial designs as one family (Games, 1971; Ryan, 1959; Stevens, 1999). For example, there are 7 hypothesis tests in the three-factor design, then, each hypothesis is tested under .05/7 using the Bonferroni procedure. The more summarized opinion was made in Miller (1981) toward the factorial design:

If the number of rows and columns is not excessive, and the degrees of freedom are adequate, the author might like to treat all statements on row
effects, column effects, and interaction effects as one family. For a large two-way design, this may not be practical. (p. 35)

Finally, Miller (1981) concluded that the real choice depends on researchers’ own purposes of design.

Multiple Hypothesis Testing (MHT) Applications

Researchers sometimes need to collect large datasets from different groups or under several conditions or with several outcome measures (Schochet, 2008). It is very common to conduct many hypothesis tests in one study (Keselman, Cribbie, & Holland, 2002). The problem of Type I error inflation in MHT would occur when a research design involves collecting data from more than two groups or under more than two conditions (Ryan, 1959). That is, if more than one statistical hypothesis is tested, researchers run the risk of inflating the Type I error. The Type I error rate inflation is the central problem in MHT (Benjamini & Yekutieli, 2001; Schochet, 2008). Therefore, researchers should be aware of the accumulative errors that result from MHT.

Many researchers are familiar with post hoc multiple comparison procedures, which are able to maintain the Type I error rate at $\alpha$ when a set of comparisons is made among sample means, following a significant one-way ANOVA. These procedures, such as the Tukey method, the Tukey/Kramer (TK) method, and Scheffé’s method, can tell where significant differences exist among groups. In fact, a great deal of literature has focused on multiple comparisons and numerous post hoc procedures are available for researchers (Stevens, 1999).
In previous literature, “multiple comparison” has been used synonymously with MHT (Shaffer, 1995). In fact, multiple comparison procedures (MCP) are common examples of MHT. As Ryan (1959) noted that MCPs mainly focus on all group mean comparisons. However, MHT has a broad scope which includes the median hypotheses, the proportion hypotheses, and correlation coefficient hypotheses. The following sections illustrate several main areas in which MHT can be conducted: multiple comparisons in one-way ANOVA, multiple tests with correlated variables, multivariate analysis of variance (MANOVA), multiple Chi-square tests in differential item functioning (DIF), multiple regression, repeated measures, and factorial ANOVA.

Multiple Comparison in One-Way ANOVA

An analysis of variance (ANOVA) is often used by researchers when comparing the means of more than two groups. If there is a significant overall effect, a multiple comparison procedure follows to clarify differences between groups. Meanwhile, the Type I error rate is controlled with the MCPs (Kromrey & La Rocca, 1995). A great deal of work has been done in the development of new MCPs in the past several decades (Barcikowski & Elliott, 1996). Kirk (1995) introduced 22 kinds of MCPs that are used by researchers. The comparison approaches are called post hoc procedures (e.g., Field, 2005). Toothaker (1993) summarized that MCPs aim to control the Type I error rate inflation in one-way ANOVA and that the most commonly used procedures are Tukey’s HSD (Honestly Significant Difference), the Bonferroni procedure, and the Scheffé approach. Researchers have made some recommendations about the choice of MCPs. For example, the Bonferroni procedure adequately controls the Type I error rate, but provides
conservative results. The Tukey approach is powerful when testing a large number of means. When there is equal variance, the Tukey approach is a good choice. However, with unequal variance, the Games-Howell procedure performs best (Field, 2005).

**Multiple Tests with Correlated Variables**

It is very common for researchers to test multiple correlation coefficients in a dataset. Some researchers provided examples of multiple hypothesis tests with correlated variables (Keselman, Cribbie, & Holland, 2002; Olejnik, Li, Supattathum, and Huberty, 1997; Ryan, 1959). For example, if there are 10 variables in a dataset, 45 coefficients can be calculated. Accordingly, MHT procedures are used to control the inflation of the Type I error rate. The original and modified Bonferroni methods were adopted on a correlation matrix in the study of Olejnik et al. (1997) to test the significance of the correlation coefficients. As they concluded, the Type I error rate is controlled with these procedures.

**Multivariate Analysis of Variance (MANOVA)**

The Type I error rate inflation occurs in a MANOVA when a number of univariate ANOVAs are conducted simultaneously (Allison & Gorman, 1993). Researchers have used some MHT procedures to deal with this problem. For example, the Bonferroni method was recommended (Bray & Maxwell, 1982; Huberty & Morris, 1989; Tabachnick & Fidell, 2001). Green and Salkind (2004) recommended using the Bonferroni method in MANOVA in their SPSS instructional book. For example, the nominal alpha should be divided by 3 if there are three univariate ANOVAs in MANOVA.
Differential Item Functioning (DIF)

MHT not only can be used in statistical area, but also in the field of measurement. “DIF occurs when members of the two groups possessing the same proficiency have different probabilities of getting the studied item correct” (Donoghue & Allen, 1993, p. 131). When evaluating parameters such as the person’s ability ($b$) or the item discrimination ($a$) within a set of items which have DIF in the reference and focal groups, the Type I error rate inflation should be paid attention in multiple chi-square tests that are involved in detecting DIF. The B-H method, developed in Benjamini and Hochberg (1995), was used in the study of Steinberg (2001) to control the Type I error rate when evaluating a set of items.

Multiple Regression

Multiple regression involves MHT. Several researchers have focused on the Type I error inflation in stepwise regression. For example, if the $p$ value of a variable is less than .05, it will be selected from several variables and be added in the regression model. As Pohlmann (1979) stated that, “A Type I error would occur if a variable was selected, using the $F$ ratio criteria” (p. 47). In fact, the variable selection procedure is similar as conducting each hypothesis at .05 in a given family which has a set of hypotheses. The Type I error rate will apparently inflate. Other researchers also emphasized the problem of Type I error inflation in stepwise regression analysis (Beasley & Leitner, 1994; Cramer, 1972; Korn & Graubard, 1990).

Besides the stepwise regression, when testing the significance of each coefficient at one time, the Type I error rate inflation should also be controlled (Mendenhall &
Sincich, 1996). They stated that, “Just as with *t* tests on individual beta parameters, you should avoid conducting too many partial *F* tests. Regardless of the type of test (*t* test or *F* test), the more tests that are performed, the higher the overall type I error rate will be” (p. 338). They also recommended limiting the number of models that are proposed.

**Repeated Measures**

In the design of repeated measure, multiple hypothesis tests are conducted among several repeated trials simultaneously. Therefore, MHT procedures should be used to control the Type I error inflation. The Bonferroni procedure was recommended in the repeated measure design to control the Type I error rate by researchers (Keselman, Keselman, & Shaffer, 1991; Maxwell, 1980; Mayers, 1979). Barcikowski and Elliott (1996) suggested that pairwise multiple comparison procedures be used in a single group repeated measure design. Looney and Stanley (1989) conducted exploratory repeated measure analysis for two or more groups and the adjusted-alpha approach was used in testing group effects and trial effects. They concluded that the adjusted-alpha method should be used when both univariate ANOVA and MANOVA are performed with repeated measures. Robey and Barcikowski (1986) investigated the Type I error in the single group repeated measures design with multiple measures. They also recommended using an adjusted-alpha approach.

**Factorial Analysis of Variance (ANOVA)**

The Type I error rate inflates in factorial ANOVA because several *F* tests are usually conducted in the same factorial design simultaneously. Halderson and Glasnapp (1974) emphasized that the multiplicity problem exists not only in one-way ANOVA, but
also in higher order designs. Researchers have paid attention to the Type I error rate inflation in factorial designs and adopted some MHT procedures to control this kind of problem. The Bonferroni method was recommended in a factorial ANOVA design to control the Type I error rate by some researchers (Rosenthal & Rubin, 1984; Smith, Levine, Lachlan, & Fediuk; 2002; Stevens, 1996). Kromrey and Dickinson (1995) found that the Hochberg method (1988) performs slightly better than other approaches (the unprotected and protected $F$ test, the Bonferroni method, and the Holm procedure) in terms of power.

**Multiple Comparison Procedures in One-way ANOVA**

Cribbie and Keselman (2003) talked about the idea of simultaneous MCP which “conduct all comparisons regardless of whether the omnibus test, or any other comparison, is significant (or not significant) at a constant critical value” (p. 169). The classical Bonferroni procedure, the Tukey HSD/ Tukey-Kramer procedure, the Dunn-Šidák procedure, and the Scheffé procedure are all regarded as examples of simultaneous MCPs (Toothaker, 1993). Meanwhile, these procedures are commonly used when the design has several treatment and control groups (Schochet, 2008; Toothaker, 1993).

**The Tukey HSD and Tukey/Kramer Method**

Tukey’s HSD (Honestly Significant Difference) method is conservative, which maintains the overall alpha ($\alpha$) level at the priori $\alpha$ level by using the studentized range ($Q$) distribution (Hinkle, Wiersma, & Jurs, 2003; Kirk, 1995; Stevens, 1999). For equal sample size, Tukey HSD is preferred; for unequal sample size, Tukey/Kramer should be used. Stevens (1999) listed three advantages of the Tukey procedure. First, it does control
the overall $\alpha$ as it supposed to. Second, it examines all paired comparisons. Third, it is a fairly powerful procedure to identify differences. He concluded that “the Tukey procedure provides a nice balance in terms of controlling on both type I and type II errors” (p. 86). The Tukey procedure is the most popular MCPs; however, it is appropriate only when all pairwise comparisons are performed (Toothaker, 1993).

*The Dunn-Šidák Procedure*

In fact, Dunn (1961) applied the Bonferroni procedure to MCP. The Dunn’s procedure provides an upper bound to the familywise Type I error rate (Kirk, 1995). Based on Dunn’s work, Šidák (1967) proposed a slightly less conservative approach by a multiplicative inequality. In this method, each hypothesis is tested at $1 - (1 - \alpha)^{1/n}$, rather than $\alpha/n$ in the Dunn-Bonferroni procedure. Therefore, this procedure is a bit more powerful than the Bonferroni procedure. However, Schochet (2008) noted that the Sidak procedure could not control the FWER in situations when tests are dependent or correlated by reviewing previous literature.

*The Scheffé Procedure*

The Scheffé approach is based on the $F$ distribution (Ryan, 1959). Hancock and Klockars (1996) stated that “Scheffé’s test is an infinite intersection test, yielding a critical value that controls the familywise error rate over all possible post hoc contrasts to exactly the nominal $\alpha$ level” (p. 298). Unlike Tukey’s method, which is suitable only for pairwise comparison, Scheffé’s method could be applied in all possible comparisons: pairwise, nonpairwise, orthogonal polynomials. It is a versatile method which can test complex hypotheses (Hinkle, Wiersma, & Jurs, 2003). The flexibility is the advantage of
Scheffé’s procedure (Stevens, 1999). Infinite numbers of hypotheses could be included in this method (Toothaker, 1993). If the number of contrasts is more than 20, the Scheffé method is more powerful than the Bonferroni procedure, but less powerful than the Bonferroni for seven and fewer tests (Schochet, 2008; Toothaker, 1993). It is less powerful than Tukey’s method and may work best only for nonpairwise contrasts (Kirk, 1995). Therefore, it is the most conservative of the MCPs and is recommended only for complex contrasts (Hinkle, Wiersma, & Jurs, 2003). In addition, Scheffé’s result is always consistent with the overall $F$ test. That is, if the $F$ test is significant, there is at least one significant result in Scheffé’s procedure (Kirk, 1995; Schochet, 2008).

Procedures in Multiple Hypothesis Testing

As Hancock and Klockars (1996) stated “New methods have been created in the continued quest to bring the empirical alpha level closer to the nominal alpha level – that is, the quest to maximize the experimental power while control over the familywise error is maintained” (p. 270). Many MHT procedures have been derived in order to control the Type I error with certain statistical power. Some researchers suggested that the adjusted $p$ values should be used for simultaneous inference (Shaffer, 1995; Toothaker, 1993; Westfall & Young, 1993; Wright, 1992). Both the unadjusted alpha per test approach and several adjusted-alpha procedures will be reviewed in the following sections.

*Unadjusted Alpha Per Test Approach*

Traditionally, hypothesis testing is often conducted at a specified alpha level. The most frequently used levels of significance are .01 and .05 (Hinkle, Wiersma, & Jurs, 2003). That is, researchers are aware that there is a 1% or 5% chance that they are
making Type I errors. For example, if .05 is set, all $p$ values (e.g., from computer output) produced in one statistical analysis should be compared to .05 and final conclusions are made. However, the Type I error rate inflation occurs when a set of hypotheses are tested simultaneously, if each hypothesis test is compared to .05. A formula is used to indicate the probability of making at least one Type I error when multiple null hypothesis are true: $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .05 and $k$ is the number of hypothesis tests (Hochberg & Tamhane, 1987; Maxwell & Delaney, 2000; Schochet, 2008; Stevens, 1996; Toothaker, 1993).

The Bonferroni Procedure

Bonferroni’s study, as cited in Hancock and Klockars (1996), derived the Bonferroni Inequality from a probabilistic proof. In fact, the derivation of the Bonferroni Inequality was not in the context of MCPs (Shaffer, 1995). Dunn (1961) applied this kind of procedure to multiple comparisons (cited in Hancock & Klockars, 1996). In the Bonferroni procedure, the alpha ($\alpha$) level is divided by the number of hypothesis tests, then the overall familywise Type I error remains below .05.

Some MHT procedures only control the Type I error rate when there are all true null hypothesis tests, other procedures work when there are some true null hypotheses and some false null hypotheses. The former situation is referred as weak control and the latter as strong control (Shaffer, 1995; Hochberg & Tamhane, 1987). The Bonferroni correction has become a standard and well-known approach for controlling the FWER in a strong sense (Cai, 2006; Shaffer, 1995; Schochet, 2008). It also serves as the foundation for many developments and improvements in MHT procedures (Hancock & Klockars,
1996). It is a flexible method because it can be applied to test any subset of hypotheses, both continuous and discrete data, and even correlated tests (Schochet, 2008). It is a simple but general procedure that does not require any constraining assumptions (Hochberg & Benjamini, 1990). This procedure “has few competitors in a variety of settings” (Holland & Copenhaver, 1987, p. 417). However, this procedure is conservative because the adjusted alpha level becomes very small if there are numbers of hypothesis tests (Field, 2005; Games, 1971; Hancock & Klockars, 1996; Holm, 1979). As Field (2005) stated, “by being more conservative in the Type I error rate for each comparison, we increase the chance that we will miss a genuine difference in the data” (p. 340). Some researchers presented that the Bonferroni procedure has low power because of using conservative critical values (e.g., Hochberg & Benjamini, 1990; Schochet, 2008).

**The Modified Bonferroni Procedures**

Since a study by Games (1971), a number of modifications and improvements have been made based on the Bonferroni procedure. Unlike simultaneously conducting a multiple hypothesis test, sequential or stepwise strategies are emphasized by researchers (Hancock & Klockars, 1996). Toothaker (1993) stated that “Stepwise methods are those that do the tests by employing some sequence of steps, each depending on the ones before it” (p. 35). However, sequential methods are not able to have simple confidence intervals, but simultaneous procedures can (Hancock & Klockars, 1995; Holland & Copenhaver, 1987; Toothaker, 1993).

**The Holm Procedure**
Holm (1979) was the first to formally introduce a sequentially rejective Bonferroni procedure. It is also called a “step-down” procedure in which the hypotheses are tested based on the ordered \( p \) values from the smallest to the largest (Kromrey & Dickinson, 1995). Compared to the classical Bonferroni procedure, Holm’s approach is less conservative and more powerful because the adjusted \( p \) values in this method sequentially increase while Bonferroni’s critical \( p \) value is fixed (i.e., \( \alpha/n \)). This method is also able to control the Type I error rate in a strong sense (Schochet, 2008). The Holm procedure is currently the most powerful procedure for controlling the Type I error rate among procedures that do not need strong assumptions, such as independence (Ge, Sealfon, Tseng, and Speed, 2007).

The Hochberg Procedure

Hochberg (1988) proposed another sequentially rejected Bonferroni procedure which is called “step-up” procedure. This method is proved to be powerful than the Holm approach based on a Monte Carlo simulation in factorial designs (Kromrey & Dickinson, 1995). However, it lacks the stability under certain conditions, for example, when the test statistics are dependent or correlated (Schochet, 2008).

Admittedly, a number of researchers attempted to make improvements on the traditional Bonferroni method, such as Hommel (1988), Shaffer (1986), Sim (1986), Rom (1990). All these approaches do improve the power slightly compared to the traditional Bonferroni, however, they are all at the cost of complexity (Hancock & Klockars, 1996; Shaffer, 1995).
The B-H Procedure

Benjamini and Hochberg (1995) described the concept of false discovery rate (FDR), which is a different perspective from FWER to control Type I error rate. In this method, FDR is the expected proportion of false rejections among all rejections. Benjamini and Hochberg (1995) provided some justifications for controlling FDR rather than FWER. First, for example, in exploratory research, a few errors of inference should be tolerable; thus, one can use the less stringent FDR method of control. Second, when two treatments are compared in multiple subgroups, the primary interest is the treatment effects within each subgroup rather than the overall effect. The motivation to use FDR was reviewed in Shaffer (1995). As she explained, researchers usually have a common misconception about the overall error rate $\alpha$ when applied to a family of tests. That is, “on the average, only a proportion $\alpha$ of the rejected hypotheses are true ones, i.e. are falsely rejected” (p. 567). If all the hypotheses are true, then it is possible to have 100% false rejection. This situation would suggest that researchers need to consider the proportion of false rejection in other points of view.

As Benjamini and Hochberg (1995) proved, there are several important properties behind this error rate (see Table 1). First, when there are all null hypotheses, the number of true null ($m_0$) equals to the number of total hypothesis ($m$), therefore; FDR actually is FWER, therefore, “control of the FDR implies control of the FWER in the weak sense” (p. 291). When the number of true null ($m_0$) is smaller than the number of total hypothesis ($m$), the FDR is smaller than or equal to the FWER. That is, controlling FWER is also controlling FDR. Therefore, the goal of controlling the Type I error rate is
achieved. Second, they proposed that controlling FDR is more powerful than FWER in MHT when there are more nonnull hypotheses. In particular, FDR is more powerful than FWER in experiments with many treatment groups (Keselman, Cribbie, & Holland, 2002; Schochet, 2008). William, Jones, and Tukey (1999) proved that FDR yields greater power than the widely used Bonferroni technique. Meanwhile, as Benjamini and Hochberg (1995) stated, “for independent test statistics and for any configuration of false null hypotheses, the procedure controls the FDR at $q^*$” (p. 293). The $q^*$ is set to equal $\alpha$ so that the FDR is controlled at $\alpha$, usually .05 because $E(Q) \leq \frac{m_0}{m} q^* \leq q^*$. In addition, this procedure not only works well for independent statistics, but also controls FDR for positively correlated normally distributed test statistics (Benjamini & Yekutieli, 2001). Thissen, Steinberg, and Kuang (2002) proved that the FDR is very simple to implement with spreadsheet. Currently, in the handbook of the National Assessment of Educational Progress (2009), the FDR procedure is said to be more suitable than other procedures and becomes the standard method to deal with the MHT situations. Keselman, Cribbie and Holland (2002) found that the B-H method is powerful especially when testing the nonnull effects and consequently, recommended using the B-H method to control the FDR when there are large numbers of tests in the family.

**Modified FDR Procedures**

Benjamini and Hochberg (2000) developed an adaptive B-H method if $m_0$ is unknown. Although this adaptive method has been proved to be more powerful than the original B-H, it lacks the theoretical foundation (Kwong, Holland & Cheung; 2001). There are several other procedures which focus on the FDR under different conditions.
Benjamini and Liu (1999) provided a step-down modified B-H method to control FDR under independent test statistics. Benjamini and Yekutieli (2001) investigated the control of the FDR in multiple testing under dependency, especially when test statistics are negatively correlated. Troendle (2000) derived two more powerful methods under the normality of test statistics. Kwong, Holland and Cheung (2002) proposed another modified B-H method to control the FDR under the joint distribution of test statistics is known and continuous. Storey (2002) found a direct approach to FDR under a particular Bayesian scenario. Admittedly, these methods are very specific but without much added benefit. Therefore, the details will not be discussed in this review.

Several Issues in Factorial ANOVA

Factorial designs have been commonly used in the educational and psychological area (Vallejo, Ato, Fernandez, & Livacic-Rojas, 2008). “A factorial design is one in which all possible combinations of the levels of two or more treatments occur together in the design” (Kirk, 1995, p. 364). The factorial design not only enables a researcher to examine the joint (interactive) effects, but also make the efficient use of resources because, for example, in a two-way Analysis of Variance (ANOVA), all subjects are tested simultaneously and the sample sizes required are half as that are needed for two one-way ANOVAs (Hinkle, Wiersma, & Jurs, 2003; Kirk, 1995; Stevens, 1999; Maxwell & Delaney, 2000). However, it is a complex issue in terms of interpreting the interaction effect in factorial designs (Kirk, 1995).
**Balanced and Unbalanced Design**

In factorial designs, if each cell has the same number of participants, it is called a balanced or orthogonal design (Stevens, 1999). Equal cell sample size in factorial designs is often used to make the calculations of sum of squares easy, although it is not an assumption (Maxwell & Delaney, 2000). In addition, in the balanced design, each effect is uncorrelated, and with relatively large $N$, the $F$ ratio of each effect is uncorrelated. That is, each effect is orthogonal to each other, which makes the interpretation of results clean and clear (Stevens, 1999). However, in reality, things often become complex and it is very possible for researchers to collect data with disproportional cell sizes. That is, when the sample size varies from cell to cell, it is called an unbalanced or nonorthogonal design because each effect is no longer orthogonal (Maxwell & Delaney, 2000; Stevens, 1999). Maxwell and Delaney (2000) also showed that it is very typical to have a nonorthogonal design in practice. Stevens (1996) discussed the proportional and disproportional designs. For example, the sample sizes per cell in the two-way balanced design are 20. In another design, the sample sizes per cell are 20, 40, 20, and 40. Although the numbers are varied from cell to cell, it is still proportional to the balanced design of 20 per cell. If the sample sizes per cell are 20, 30, 40, and 50, it is actually a disproportional design. Therefore, the unbalanced design requires that the sample sizes per cell are not only unequal, but also disproportional to each other.

Researchers often have unbalanced designs in ANOVA (Maxwell & Delaney, 2000; Rakow, 1995). The results are not precise as expected in the unbalanced design (Rakow, 1995). Therefore, researchers should be very careful when dealing with the
unbalanced design because different options in the statistical packages would lead to conflicting results (Rakow, 1995). For the balanced design, it is very possible for researchers to make a wise, clear, and complete explanation about the outcomes. However, the interpretation becomes difficult in the unbalanced design (Vallejo, Ato, Fernandez, & Livacic-Rojas, 2008). Therefore, Keppel (1991) suggested that unbalanced designs should be avoided.

Consequences of Assumption Violation in Factorial Design

It is important to study the assumptions underlying ANOVA (Stevens, 1996), that is, “Because in ANOVA, we set up a mathematical model based on the assumptions, and all mathematical models are approximations to reality. Therefore, violations of the assumptions are inevitable” (p. 75). Researchers have focused on the assumption violation in factorial one-way and two-way ANOVA for several decades (Glass, Peckman, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992). The results are the same in factorial two-way ANOVA as that in one-way under assumption violations (Harwell, Rubinstein, Hayes, & Olds, 1992).

There are some assumptions underlying the ANOVA (Field, 2005; Hinkle, Wiersma, & Jurs, 2003; Maxwell & Delaney, 2000; Stevens, 1996; Tabachnick & Fidell, 2001): First, the participants are randomly selected from normally distributed populations, or randomly assigned to the treatment levels. Second, both the numerator and denominator of the $F$ ratio are independent and share a common variance (the population variance in all cells of the factorial design are equal). Third, the scores are independent of each other. However, in real experiments, assumptions are not always
The consequences of violating the assumptions in factorial designs have been investigated by researchers. For example, Glass, Peckman, and Sanders (1972) concluded that the assumption violation has brought considerable difficulties in analyzing data and lead to great concerns about the validity of the statistical analysis. Stevens (1996) stated that “all violations are not serious; the important thing is to know which assumptions to be particularly concerned about under what conditions” (p. 237).

**Independence**

Independence affects the Type I error rate and statistical power in $F$ test (Harwell, 1991). Glass, Peckman, and Sanders (1972) also emphasized the serious effects of the non-independence on the validity of $p$ values in the $t$ tests or ANOVA. That is, the higher the correlations between the hypotheses tests, the more deviation from the nominal alpha. Generally, Glass, Peckman, and Sanders (1972) stated that “Non-independence of scores seriously affects both the level of significance and power of the $F$ test regardless whether $n$’s are equal or unequal” (p. 273).

**Normality**

There is little effect of the non-normality on the two-tailed $t$ test and $F$ test (Glass, Peckman, & Sanders, 1972). That is, $F$ test is robust with respect to the non-normality violation (Kirk, 1995; Stevens, 1996). Moderate non-normality does affect power, and negative Kurtosis tends to reduce power (Harwell, Rubinstein, Hayes, & Olds, 1992). Other researchers found that skewness has little effect on the Type I error rate and statistical power in fixed-effects ANOVA (Glass, Peckman, & Sanders, 1972; Kirk, 1995; Stevens, 1996). As for Kurtosis, it does have a small influence on the alpha but a
considerable effect on the power of the $F$ test (Glass, Peckman, & Sanders, 1972). Stevens (1996) concluded that platykurtosis affects power, in particular, attenuates power. Maxwell and Delaney (2000) also stated that “ANOVA is generally robust to violations of the normality assumption; in that even when the data are nonnormal, the actual type I error rate is usually close to the nominal value” (p. 109).

**Equal Variance**

Box’s study, as cited in Kirk (1995), concluded that the $F$ test is robust under the violation of homogeneity of variance assumption under certain conditions, such as, equal cell size, normally distributed populations, and the ratio of the largest to the smallest variance is not over 3.0. However, Kirk (1995) also suggested that the violations of homogeneity of variance assumption not be ignored. Stevens (1996) illustrated that ANOVA is robust under unequal variances if group sizes are equal or approximately equal (largest/smallest<1.5). The parametric analysis in comparing group means is very sensitive to the unequal population variance when the sample sizes are obviously unequal (Olejnik and Algina, 1985). When the group sizes are substantially unequal with unequal variance, $F$ test is not robust. For example, Stevens (1999) stated that “if the large sample variances are associated with the small group sizes, the $F$ statistics is liberal” (p. 249), in addition, “When the large variances are associated with the large group sizes, then the $F$ statistics is conservative” (p. 250). Unequal variance with equal sample size in the fixed-effects ANOVA has very slight effect on alpha: The actual alpha ($\alpha$) is slightly larger than the nominal. However, if unequal variance combining with unequal sample size, alpha ($\alpha$) may be seriously affected. For example, the actual ($\alpha$) tends to be greatly
larger than the nominal ($\alpha$) when smaller samples are drawn from more heterogeneous populations, the reverse is true (Glass, Peckman, & Sanders, 1972).

*Statistical Power Analysis in Factorial ANOVA*

Cohen (1988) explicitly illustrated the power analysis for different statistical designs with detailed tables. Kline (2004) noted that the power analysis for different designs including factorial ANOVA can be done with several computer packages (e.g., PASS, GPOWER, and MC4G). The factorial ANOVA is more powerful than conducting several one-way ANOVAs at one time. “A second advantage of factorial designs is that they can lead to more powerful tests by reducing error (within cell) variance” (Stevens, 1999, p.148).

Unlike one-way ANOVA which only has one overall effect, there are main effects and interaction effects in factorial ANOVA, therefore, the power determination involves considering all effects. Some researchers investigated the power in the balanced factorial ANOVA. Levin (1997) analyzed the power for the fixed-effects factorial ANOVA. For a certain pattern of cell mean in a factorial design, the effect size in the row effect, column effect, and interaction effects are different. Therefore, the power in each effect is different. The author suggested that effect size in each main effect and interaction effect should be estimated respectively, and accordingly, the determination of appropriate sample size can provide adequate power in each effect. Neter et al. (1996) suggested that the sample size in each main effect should be adequately determined in multifactor studies so as to ensure enough power of the design. Example of power analysis was illustrated in Stevens (1999) for the two-way and three-way ANOVA. As Stevens (1999)
stated, “the main reason for setting up a factorial design is to test for an interaction effect. Unfortunately, the power for detecting this interaction can be inadequate” (p. 188). The reasons for inadequate power in the example are relative small sample size per cell and small effect size (Stevens, 1999). Smith et al. (2002) investigated both Type I error rates and Type II error rates in factorial ANOVA with Monte Carlo simulation. In their study, the power rate was very low with small sample size per cell (e.g. \( n = 20 \)) and small effect size. With the sample size per cell increases (\( n = 200 \)), the power increased greatly. With very large effect size and sample size per cell, the Type II error did not exist. As they concluded, “the lesson here is that unless researchers have good reason to expect large effects, much larger samples are frequently needed” (p. 526).

Several MHT procedures are applied in factorial ANOVA in this study. The concept of power of in the context of multiple comparisons is also applied in MHT (Kromrey & Dickinson, 1995). In the context of factorial ANOVA, the power is obtained based on all hypothesis tests of the main effects and interaction effects. There are specific definitions about the power in Kromrey and Dickinson (1995): The probability of finding at least one true significant effect within the main and interaction effects is called any-effects power. The probability of finding all true significant effects is all-effects power. If averaging the per-pair power, it is called average per-effect power. Some operational definitions of power indices are provided in Chapter three.

Several Related Monte Carlo Studies Mainly in Factorial ANOVA

Several researchers have paid attention to the Type I error rate inflation in multifactor design. In particular, Monte Carlo techniques were employed to investigate
the Type I error rate and statistical power rate in these research. Since the current study will make some extensions based on these related articles, the following are brief reviews of these Monte Carlo studies.

In the empirical study of Halderson and Glasnapp (1974), three kinds of error rates (the per comparison, per experiment, and experimentwise error rates) were investigated for three hypothesis testing procedures (the alpha procedure, the Hartley sequential procedure, and the Bonferroni procedure) under certain varied conditions in factorial ANOVA. Both balanced two-way and three-way fixed-effects designs were investigated. More specifically, 2x2, 2x3, 2x4, 2x5, 5x5, 2x2x2, and 5x5x5 designs were generated. The per cell observation was 5, 15, and 30. The effect size was mainly based on Cohen’s standard: small ($f = .10$), medium ($f = .25$), and large ($f = .40$). All data were normally distributed and with equal variance. The Monte Carlo study had 2000 replications.

In their results, the experimentwise error rates were close to .15 and .30 which is close to the formula $1 - (1 - \alpha)^k$ where $k$ is the number of hypothesis test. Both the Bonferroni procedure and the Hartley procedure controlled the experimentwise error rate at acceptable levels. Finally, the authors concluded that the choice of a MHT procedure depends on the importance of errors. If individual test is a focal point, the alpha procedure can be used. If the entire design is focused on, either the Bonferroni procedure or the Hartley procedure can be used.

The study of Fletcher, Daw, and Young (1989) also confirmed the problem of multiple $F$ tests in factorial designs. Accordingly, a Monte Carlo simulation was
conducted with a three-factor balanced ANOVA (2x2x2). The per cell observation was 5 and the iteration was 100. The overall $F$ test and the Bonferroni procedure were evaluated in terms of the Type I error rate. As a result, the Type I error rate of 32% in the three-way balanced design was obtained which was consistent with the formula $1 - (1 - \alpha)^k$ where $k$ is the number of hypothesis test. Meanwhile, the Bonferroni procedure had more error rate than the overall $F$ test and accordingly, the latter was recommended by the authors based on their study.

The study by Kromrey and Dickinson (1995) overcame the limitations in Fletcher, Daw, and Young (1989) and provided more comprehensive solutions by investigating the Type I error rates and statistical power rate in factorial ANOVA. Kromrey and Dickinson (1995) pointed out that the overall $F$ test which Fletcher et al. (1989) recommended only works when all null hypotheses are true. In addition, the 100 replications were insufficient to conclude that the Bonferroni procedure has greater Type I error rate than the overall $F$ test in Fletcher et al. (1989).

A Monte Carlo simulation in factorial ANOVA was conducted in Kromrey and Dickinson (1995). The Type I error rates and statistical power rates were investigated. Five procedures were used: a) unprotected $F$ test, b) protected $F$ test, c) the Bonferroni procedure, d) the Holm procedure, and e) the Hochberg procedure. 2x2, 2x2x2, and 2x2x2x2 balanced designs were included, each factor only has two levels. The per cell observation was 5, 10, and 20. The patterns of true null hypothesis and false null were varied. The effect size was based on the standard in Cohen (1988). There were 5000 iterations. All the data generated were normally distributed and with equal variance.
As a result, the Bonferroni procedure and the modified Bonferroni procedure (the Holm procedure and the Hochberg procedure) controlled the familywise Type I error rates better than the $F$ test under partial null conditions. The modified procedures were more relatively powerful than the original Bonferroni procedure. In particular, the Hochberg procedure was the most powerful. Therefore, the Hochberg procedure was recommended in term of relative power advantage.

A Monte Carlo simulation was conducted in the study of Smith, Levine, Lachlan, and Fediuk (2002). Three-way and four-way balanced factorial designs were investigated. The cell sizes were 20 or 200. The number of effects was specified as 0, 1, or 3. Small or medium effect size was used. The Bonferroni procedure and the overall $F$ test were included. There were 500 trials in the simulation. Finally, the Bonferroni procedure was recommended because of its stability in controlling the Type I error rate under both complete null and partial null situations.

Schochet (2008) investigated the statistical power rate of four procedures: no adjustment, the Bonferroni procedure, the Holm procedure, and the B-H procedure. Multiple $t$ tests were generated independently. The number of $t$ tests was 5, 10, 20, and 30. Each simulation had 10000 replications. The percentage of true null hypothesis was 80, 50, and 20. As a result, the power losses in the B-H procedure were smaller than that in the Bonferroni and Holm procedure.

Chapter Summary

Type I error rate and statistical power are important indicators in terms of evaluating MHT procedures. Researchers should be aware that what kind of Type I error
rate is to be controlled and what kind of power is to be reported. Whether adjust alpha or not is a philosophical issue and there are different opinion toward this question. Some researchers suggested that the alpha not be adjusted because of power loss and inconsistent results. That is, the unadjusted procedure should be always used in any situation (O’Keefe, 2003). Some researchers proposed that alpha-adjustment procedures should be used in appropriate contexts (Hewes, 2003; Tutzauer, 2003). Many researchers argued that MHT procedures should be used when many hypotheses are tested simultaneously. However, the solution is not unique because of different definitions of family.

MCPs are one example in MHT (Ryan, 1959; Schochet, 2008; Shaffer, 1995). The review and discussions about MHT procedures also include the MCPs, therefore, some MCPs in one-way ANOVA were reviewed in this Chapter, such that the Tukey procedure, the Dunn-Šidák procedure, and the Scheffé procedure. These procedures can control the FWER in one-way ANOVA. Researchers made some recommendations when using these MCPs (Field, 2005; Schochet, 2008; Stevens, 1996; Toothaker, 1993). If all pairwise comparisons are the primary interests in the study, the Tukey procedure is recommended. If all possible contrasts are being considered, the Scheffé procedure is recommended. The Dunn-Šidák procedure is similar to the Bonferroni method.

Some MHT procedures that were reviewed in this section are mainly controlling FWER, such as the Bonferroni method, the Holm method, and the Hochberg method. Both the Bonferroni procedure and the Holm procedure are recommended as suitable general-purpose methods in MHT (Holland & Copenhaver, 1987; Schochet, 2008)
because of no constraining assumptions. The Holm procedure is proved to be more powerful than the Bonferroni (Holm, 1979). The explanation of the Bonferroni is easy (Schochet, 2008). The Hochberg procedure is more powerful than the Bonferroni method and the Holm procedure (Hochberg, 1988). In the Monte Carlo study in Kromrey and Dickinson (1995), the Hochberg procedure shows more relative power than other procedures (the unprotected and protected $F$ tests, the Bonferroni method, and the Holm procedure). However, this procedure can control the FWER only when the test statistics are independent (Holland & Copenhaver, 1987).

The B-H method focuses on controlling FDR. Schochet (2008) stated that “this ‘step-up’ sequential procedure, which has become increasingly popular in the literature, is easy to use because it is based solely on $p$ values from the individual tests” (p. 18). It has been proved to be more powerful than some FWER methods (e.g., the Bonferroni and the Holm procedure) (Keselman, Cribbie, & Holland, 2002; Schochet, 2008). Williams, Jones and Tukey (1999) proved that the B-H method is more powerful than the Bonferroni procedure and the Hochberg method.

FDR has a different philosophical rationale from FWER (Benjamini-Hochberg, 1995; Schochet, 2008). FWER focuses on any false rejection of true null hypothesis. That is, a few false rejections may be tolerable and not problematic in a family with many hypotheses tests (Benjamini-Hochberg, 1995. The choice of controlling FWER and FDR is important in the research designs (Schochet, 2008). As Benjamini and Hochberg (1995) showed, on one hand, FDR controls FWER in a weak sense when there are all true null hypotheses; on the other hand, any procedure controls FWER also controls FDR.
However, few studies have focused on the connections between FWER and FDR from the practical point of view.

Factorial ANOVA has become common and useful techniques that are available for researchers. This kind of design shows some advantages than one-way ANOVAs. Theoretically, the balanced designs are easier to be analyzed and interpreted than the unbalanced designs. However, researchers often encounter the unbalanced designs in practice. Meanwhile, there are certain assumptions underlying the factorial designs. The consequences of violating assumptions were also discussed in this chapter.

The accumulated Type I error rates in factorial ANOVA have been investigated by some researchers using Monte Carlo simulation. The review of “Closest” articles provided general ideas about how these Monte Carlo studies were conducted in factorial designs. However, not many researchers have investigated the Type I error rates and statistical power with several FWER and FDR procedures in one factorial ANOVA. Therefore, a Monte Carlo study was illustrated in Chapter three to investigate the Type I error rates and statistical power in the two-way and three-way factorial ANOVA.
CHAPTER THREE: METHODOLOGY

A mathematical analysis is not workable under the following conditions: (a) statistical assumptions are violated, (b) conditions that the mathematical theories can be applied are not met (e.g., the null hypothesis is known not to be true), and (c) a statistic for the sampling distribution has not been solved in a mathematical way (Mooney, 1997).

In the area of statistical inference, because statistical procedures are performed under certain strict assumptions, researchers often have concern about the validity of statistical inference when those assumptions are violated. Monte Carlo studies enable researchers to identify how sensitive the tests are to assumption violations. Meanwhile, some true null hypotheses and false null hypotheses exist together in one design, Monte Carlo method could help researchers to know the true null and false null hypotheses, and then the Type I error rates and statistical power rates can be obtained (Harwell, Rubinstein, Hayes & Olds, 1992).

The primary purpose of this study was to investigate how well several MHT procedures work in terms of the Type I error rates and statistical power rates in the balanced two-way and three-way factorial ANOVA. Three criteria were used to indicate the Type I error rates of MHT procedures. These criteria were Samplewise FWER (SFWER), Testwise FWER (TFWER), and false discovery rate (FDR). Meanwhile, three indicators were used to evaluate the statistical power rates: Samplewise Power (SPOWER), Testwise Power (TPOWER), and true discovery rate (TDR). To obtain these criteria, it is very necessary to generate different patterns of true null and false null hypotheses and individual samples. Replication and looping are needed. In this sense,
Monte Carlo simulation is an appropriate and reasonable technique that is employed in this study.

**Monte Carlo Simulation**

Mooney (1997) stated that “Monte Carlo simulation offers an alternative to analytical mathematics for understanding a statistic’s sampling distribution and evaluating its behavior in random samples” (p. 2). The Monte Carlo method is based on repeated random sampling to compute results (Brooks, 1998). Lix, Keselman, and Keselman stated that “In a typical simulation study, pseudorandom sets of numbers are generated using a predefined computer algorithm and are sampled from populations with known characteristics” (p. 589). Monte Carlo simulation has some applications to understand statistical and social processes: (a) Evaluating the robustness of statistical procedures under the assumption violations. (b) Testing the null hypotheses under some possible conditions. (c) Assessing the quality of inferential methods, and comparing two or more estimators (Mooney, 1997). For example, the researcher could use simulation to investigate the Type I error rates of statistical procedures over many samples, which are generated under the true null hypothesis while violating assumptions. Rejections could be counted with many iterations and the Type I error rate under the true null would be obtained. Then, researchers could compare the Type I error rates of these statistical procedures.

Mooney (1997) explained the principle behind Monte Carlo simulation. That is, “a statistic in random samples can be assessed by the empirical process of actually drawing lots of random samples and observing this behavior” (p. 3). Based on sample
parameters, the pseudo-population is generated to resemble the real population. The following steps for a basic Monte Carlo simulation were described in Mooney (1997):

1. Specify the pseudo-population in symbolic terms in such a way that it can be used to generate samples. This usually means developing a computer algorithm to generate data in a specified manner.

2. Sample from the pseudo-population (a pseudo-sample) in ways reflective of the statistical situation of interest, for example, with the same sampling strategy, sample size, and so forth.

3. Calculate \( \hat{\theta} \) in the pseudo-sample and store it in a vector, \( \mathbf{\theta} \).

4. Repeat Steps 2 and 3 \( t \) times, where \( t \) is the number of trials.

5. Construct a relative frequency distribution of the resulting \( \hat{\theta} \) values, which is the Monte Carlo estimate of the sampling distribution of \( \hat{\theta} \) under the conditions specifically by the pseudo-population and the sampling procedures.

(p. 4)

It should be noted that \( \hat{\theta} \) is the estimated parameter and \( \theta \) is the population parameter. Brooks (1998) pointed out that the Monte Carlo design is not very different from the standard research design, which generally includes (a) identification of the population, (b) description of the sampling plan, (c) instrumentation, (d) data collection, and (e) data analysis. The following sections will describe these elements of the Monte Carlo design.
Research Design

A Monte Carlo analysis in the Type I errors and statistical power of several MHT procedures in the balanced two-way and three-way factorial ANOVA were performed. Specifically, three criteria for the Type I error rates were used: Samplewise FWER (SFWER), Testwise FWER (TFWER), and FDR. Similarly, three standards were used to evaluate statistical power rate of MHT procedures: Samplewise Power, Testwise Power, and TDR. Four alpha-adjustment MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure) and the unadjusted alpha per test procedure were evaluated. There were four patterns of true null and false null effects in the balanced two-way design in which the first pattern has all true null hypotheses and the fourth pattern has all false nulls. The first three patterns (1-3) were used to investigate the Type I error rates and the last three patterns (2-4) were employed to evaluate statistical power rates. Similarly, in the three-way design, the first seven patterns (1-7) were used to calculate the Type I error rates and the last seven patterns (2-8) for the statistical power rates. It is necessary to mention that the patterns of true null hypotheses and false null hypotheses that are used in the simulation are some subset of all possible patterns of true null and false null hypotheses. But since all hypotheses are regarded as one family in this study, it doesn’t matter whether which main effect or interaction effect is significant.

The research question (RQ1a) was to investigate how well different MHT procedures control Type I error rates in terms of three criteria in the balanced two-way ANOVA, which can be considered as a 4x5 design. Similarly, the research question
(RQ2a), which investigates the statistical power in the two-way balanced design, can also be regarded as a 4x5 design. There were four patterns of true null and false null and five procedures that were compared in each research question. In addition, three measures for the Type I error rates were regarded as outcome levels. The following table illustrates the patterns of true and false effect in the two-way analysis.

Table 2

<table>
<thead>
<tr>
<th>Number of True Null</th>
<th>Pattern Numbers</th>
<th>Cell Mean (1,1)</th>
<th>Cell Mean (1,2)</th>
<th>Cell Mean (2,1)</th>
<th>Cell Mean (2,2)</th>
<th>True Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>A, B, AB</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
<td>A, AB</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.50</td>
<td>1.00</td>
<td>0.00</td>
<td>0.50</td>
<td>AB</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>None</td>
</tr>
</tbody>
</table>

Note: n per cell is 32. Pattern 1-3 for Type I error rate and pattern 2-4 for statistical power.

There are three hypothesis tests in the two-way balanced design: two main effects (A and B) and one interaction effect (AB). For example, cell means in pattern one is all equal to zero and there are no standardized mean differences. Both the main effects (A and B) and the interaction term (AB) are true null hypothesis. The cell means in pattern two are 0.0, 0.5, 0.0, and 0.5. The main effect (A) and interaction (AB) are true null hypotheses; the main effect (B) is false null. The cell means in pattern three are 0.5, 1.0,
0.0, and 0.5. Only interaction term (AB) is true null. The hypotheses are all false nulls in pattern four. Since it is a balanced design, there are 32 observations per cell.

The research question (RQ1b) was to investigate how well different MHT procedures control Type I errors in the balanced three-way ANOVA, which could be considered as an 8x5 design. Similarly, the research question (RQ2b), which investigated the statistical power in the three-way balanced design, could also be regarded as an 8x5 design. There were eight patterns of true null and false null and five procedures were compared in the three-way design. Meanwhile, three measures for statistical power rates were regarded as outcome levels. The following table illustrates the patterns of true and false effect in the three-way analysis.
<table>
<thead>
<tr>
<th>Number of True Null</th>
<th>Pattern Numbers</th>
<th>Cell (1,1,1)</th>
<th>Cell (1,2,1)</th>
<th>Cell (2,1,1)</th>
<th>Cell (2,2,1)</th>
<th>Cell (1,1,2)</th>
<th>Cell (1,2,2)</th>
<th>Cell (2,1,2)</th>
<th>Cell (2,2,2)</th>
<th>True Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>all</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>A,B,C,AB,AC,ABC</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.0</td>
<td>A,C,AB,BC,ABC</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>A,AB,AC,ABC</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>B,AC,ABC</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>A,ABC</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>AC</td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0 None</td>
</tr>
</tbody>
</table>

Note: \( n \) per cell is 16. Pattern 1-7 for Type I error rate and pattern 2-7 for statistical power.
There are seven hypothesis tests in the three-way balanced design: three main effects (A, B, and C), three two-way interactions (AB, AC, and BC), and one three-way interaction (ABC). Each cell size is 16. For example, pattern three has five true null and two false null effects. The true null are A, C, AB, BC, and ABC. Pattern seven has only one true null which is AC. All false nulls are in pattern eight.

The choice of the sample size in each cell and the effect size is not entirely arbitrary. In previous Monte Carlo studies in factorial ANOVA, 5, 15, and 30 observations per cell were used in the two-way and three-way analysis in Halderson and Glasnapp (1974). There were five observations per cell in the study of Fletcher, Daw and Young (1989). The observations of 5, 10, and 20 per cell were adopted in the study of Kromrey and Dickinson (1995). The effect size of .50, which is based on the principle in Cohen (1988), is used in this study, previous studies also used the Cohen’s standard (1988) (Halderson & Glasnapp, 1974; Kromrey & Dickinson, 1995). Stevens (1996) suggested that 20 observations per cell with medium effect size can ensure enough power in factorial design. Finally, using statistical software, such that MC4G program (Brooks, 2008) and GPOWER, given 32 sample size per cell in the two-way and 16 in the three-way design, with the effect size of .50, fairly enough power of .80 is achieved.

The choice of the numbers of nonnull effects is not entirely arbitrary. In fact, all possible numbers of true null and false null are included in the two primary research questions and .50 is used as the effect size for the nonnull effects. For example, there are three hypotheses in the two-way balanced design, all possible conditions such that three true null, two true null, one true null, and zero true null are included. For the three-way
balanced design, possible numbers of true null are also considered; the true null effect ranges from zero to seven. Although the patterns are some subset of all possible true null and false null hypotheses, it does not matter because all hypotheses are regarded as one family in this study. Some researchers suggested that the number of true null or false null should be focused on rather than the specification of nonnull effect, as far as the normal theory is concerned (Sawilowsky, Blair, & Higgins, 1989; Shaffer, 1977). That is, in the pattern which has three true null and four false null effects, the number of true null and the number of false null were emphasized instead of focusing on which effect is true null and which is the false null effect. The results will not be different no matter which effect is true and which effect is false in this design. In other words, the patterns that were used in this study have generalizability.

Monte Carlo Simulation in R

A Monte Carlo program was created using the computer statistical program called R that was used to simulate data needed to obtain the appropriate Type I error rates and statistical power rates. R is a computer statistical package and programming language available as open source software (see http://www.r-project.org/). All statistical analyses were performed using built-in R functions with the requirement of some R packages installed, such that the packages of “stats”. The false rejection rates on true nulls and true rejections on false nulls of the several MHT procedures were recorded for each sample generated under different cell mean patterns. Accordingly, the Type I error rates and statistical power rates were calculated based on their indicators.
The number of iterations (i.e., 20,000) was used in this study for the purpose of the generalizability. In previous research, the replications of 2000 were used in Halderson and Glasnapp (1974). The iterations of 100 were conducted in Fletcher, Daw, and Young (1989). Kromrey and Dickinson (1995) used 5000 replications. There were 500 runs in the study of Smith, Levine, Lachlan, and Fediuk (2002). Mooney (1997) suggested that the best practice of choosing the number of iterations is the more the better. He also mentioned that 1,000 trials are common practice and the Monte Carlo simulations have greater power with more trials. Meanwhile, based on the recommendations in Robey and Barcikowski (1992) which provided good references about selecting the number of iterations, the iterations (i.e., 20,000) in this study leads to smaller standard error of results and makes a lot of practical sense.

**Operational Definitions**

The following definitions are helpful to have a comprehensive understanding of the methodology employed in the Monte Carlo study.

**P-values**

Using $p$ values to indicate statistical significance is an effective and easy way for external readers to have a better understanding about statistical inference (Hochberg & Benjamini, 1990). The factorial ANOVA analysis is conducted in R using built-in functions. As a result, $p$ values of each effect are obtained. Then, the adjusted $p$ values in MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure) are calculated in R using the function `p.adjust`. All of
adjusted $p$ values from MHT procedures are compared to .05 because it is the $p$ values that are adjusted in R rather than the alpha level.

*Type I Error*

The Type I error occurs when falsely reject the null hypothesis. That is, the calculations of Type I errors are based solely on the null hypothesis. In the present Monte Carlo study, although there are different numbers of nonnull effects for the two-way and three-way design, the Type I error rates are calculated according to all of the true null hypotheses. The adjusted $p$ values from the true nulls are compared to .05. The null hypothesis is rejected if they are less than .05. The number of rejections is kept counting with 20,000 trials, and then the samplewise FWER, Testwise FWER, and FDR are calculated. All Type I error values are stored in an array in R.

*Familywise Error Rate (FWER)*

The FWER is the probability of at least one hypothesis is incorrectly rejected in a given family (Hochberg & Tamhane, 1987; Ryan, 1959; Shaffer, 1995; Toothaker, 1993). It is common to control FWER in MHT (Hochberg & Tamhane, 1987; Hancock & Klockars, 1996; Toothaker, 1993). Two types of FWER will be examined in this study: SFWER and TFWER.

*Samplewise Familywise Error Rate (SFWER)*

Some researchers emphasized that the experimentwise error rate (EWER) will be different from the familywise error rate in multifactor design (Maxwell & Delaney, 2000; Ryan, 1959). The reason is that there are several families in one factorial experiment. As the example illustrated in Maxwell and Delaney (2000), there are two separate families.
Controlling the EWER actually requires controlling two FWER. However, in any sample, there could be multiple experimental analyses; therefore, a more general term to represent this error is Samplewise Family Error Rate. That is, SFWER is the familywise error rate that considers each sample to be a family. In this current Monte Carlo simulation, one sample dataset is drawn to represent one experiment in each replication. With 20,000 trials, there are 20,000 experiments or sample datasets. In this sense, we can call it either EWER or SFWER, but we have chosen to use the more general term SFWER in this Monte Carlo study because it more accurately reflects the per-sample nature of the analysis desired.

To calculate the SFWER, the number of Type I errors is counted for each replicated sample, and then divided by the total samples or replications (i.e., 20,000). It is necessary to indicate here that the rejections are counted once even if multiple null hypotheses are rejected. That is, what is counted is the existence of samplewise errors no matter how many Type I errors or hypotheses. For example, under the condition with two true null hypotheses and one false null hypothesis in the two-way analysis, if rejections are found either in one true null or both, the program only counts once for this single sample. If 8,000 samples are found with the Type I errors, the SFWER would be 8,000/20,000, that is, .40.

*Testwise Familywise Error Rate (TFWER)*

The rejection based on each true null hypothesis is counted every time. That is, if both true nulls have rejections, the program counts them as two. For example, under the condition with two true null hypotheses and one false null in the two-way analysis, the
total number of tests would be 40,000 (because there are two true null in one sample dataset, Type I error only occurs under the true null). If 12,000 false rejections are found when testing the 40,000 total true null hypotheses, the TFWER is 12,000/40,000, that is, .30. The main difference between the SFWER and the TFWER is that there are several tests in one sample that is replicated.

*False Discovery Rate (FDR)*

The false discovery rate (FDR) is the proportion of the false rejections and the total rejections. The number of false rejections is the same as the numerator of the TFWER. The total rejections (i.e., the denominator of the formula) actually count all rejections among all hypothesis tests, including both true nulls and false nulls. Using the same example as above, the number of false rejections is 12,000. Assume that the total number of rejections was determined to be 60,000; the FDR is 12,000/60,000, that is, .20.

*Samplewise Power (SPOWER)*

Samplewise Power is the probability of detecting at least one true significant effect on false nulls within the samples that are generated in the Monte Carlo simulation. To calculate the SPOWER, the number of at least one true rejection is counted for each replicated sample, and then divided by the total samples or replications (i.e., 20,000). It is necessary to indicate here that the true rejections are counted once even if multiple true rejections are existing. That is, what is counted is the existence of true rejections no matter how many true rejections in each sample. For example, under the condition with two false null hypotheses and one true null hypothesis in the two-way analysis, if true rejections are found either in one false null or both, the program only counts once for this
single sample. If 16,000 samples are found with true rejections, the SPOWER would be 16,000/20,000, that is, .80.

*Testwise Power (TPOWER)*

The rejection based on each false null hypothesis is counted every time. That is, if both false nulls have rejections, the program counts them as two. For example, under the condition with two false null hypotheses and one true null in the two-way analysis, the total number of tests would be 40,000 (because there are two false null in one sample dataset, statistical power are based on the false null). If 20,000 true rejections are found when testing the 40,000 total null hypotheses, the TPOWER is 20,000/40,000, that is, .50. The main difference between the SPOWER and the TPOWER is that there are several tests in one sample that is replicated.

*True Discovery Rate (TDR)*

The TDR is the proportion of the true rejections and the total rejections. The number of true rejections is the same as the numerator of the TPOWER. The total rejections (i.e., the denominator of the formula) actually count all rejections among all hypothesis tests, including both true nulls and false nulls, which is the same as that in the FDR. Using the same example as above, the number of true rejections is 48,000. Assume that the total number of rejections was determined to be 60,000; the TDR is 48,000/60,000, that is, .80.

*Sampling Plan and Data Generation*

The data that are simulated in a Monte Carlo study reflects a specified relationship among the variables (Harwell, 1992). One dependent variable and two
factors (A and B) were generated in the two-way design and one dependent variable and three factors (A, B, and C) for the three-way. Each factor has two levels; therefore, there are four cells (2x2) in the two-way analysis and eight cells (2x2x2) for the three-way. Normally distributed dependent variables were generated in each cell using the function \texttt{rnorm} in R. The cell means are shown in Table 7 and Table 8. Different cell mean patterns produces different numbers of nonnull effects. There were 32 observations per cell for the two-way balanced design and 16 observations per cell for the three-way. The standard deviation is 1.0. Either a “1” or a “2” was generated to indicate the two levels in each factor. For example, in the balanced two-way analysis, there were 32 cases with a “1” and 32 cases with a “2” in each factor. The final dataset was obtained by combining all cell data into one data frame using R functions. With 20,000 replications, we have 20,000 datasets.

\textit{Data Collection Procedure}

Each replication was saved in R. All dependent variables and factors with four conditions of nonnull effects in the balanced two-way and eight conditions with nonnull effects in the balanced three-way were all saved during program execution. For each sample, the program performed the factorial ANOVA analysis, calculated the necessary statistics and probabilities. The \(p\) values for the main effects and interactions were pulled out and saved in a matrix. The number of rows is the replications and the column is the number of tests. For example, the matrix is 20,000 * 3 for the two-way analysis because there are three effects A, B, and AB in the two-way design. For the three-way, the dimension of the matrix is 20,000 * 7. The rejections were counted based on all \(p\) values
in each column in the matrix. The criteria for the Type I error rates (SFWER, the TFWER, and the FDR) and for the statistical power (SPOWER, TPOWER, and TDR) were calculated accordingly.

For the MHT procedures, the adjusted $p$ values were calculated using R built-in functions for each sample and then the adjusted $p$ values were also saved in separate matrices. The criteria were calculated for each of the four MHT procedures (i.e., SFWER, TFWER, FDR, SPOWER, TPOWER, and TDR). More specifically, the program calculated the following information from the factorial ANOVA analysis in R:

1. $P$ values from $F$ test (Type I sum of square is used to calculate the variance in $F$ test)
2. Adjusted $p$ values from each MHT procedure
3. Total false rejections over true null hypothesis within samples
4. Total false rejections over true null hypothesis within tests
5. Total true rejections over false null hypothesis within samples
6. Total true rejections over false null hypothesis within tests
7. Total rejections for all hypotheses
8. SFWER under all MHT procedures
9. TFWER under all MHT procedures
10. FDR under all MHT procedures
11. SPOWER under all MHT procedures
12. TPOWER under all MHT procedures
13. TDR under all MHT procedures
Verification of the Data Collection Procedures

Researchers suggested that verification of the algorithms should be included in a Monte Carlo study (Bratley, Fox, & Schrage, 1987). They provided the following steps: (a) manual verification, (b) modular testing, (c) checking the results based on known formulas, (d) sensitivity testing, and (e) stress testing. The testing and verification procedures in this study are mainly based on the suggestions above.

Brooks (1998) stated that “modular testing ensures that each subroutine produces sensible output for all possible inputs” (p. 146). Therefore, modular testing was also performed in this study. For example, the normally distributed dependent variable that was generated in R was tested to ensure that they were indeed normal data. The cell mean, cell standard deviation, cell size, cell minimum, and cell maximum were calculated by the R simulation code was compared to the descriptive results in the R testing code. The datasets for each pattern of the last replication were pulled out by the R testing code. Also, the sample size per cell is set to 1000. As a result, the cell means were close to 0.0 and the cell standard deviations were close to 1.0. In addition, the cell size, cell minimum, and cell maximum are identical. Also, all datasets for each pattern of the last replication were analyzed in SPSS (SPSS Inc., 2008). The descriptive results in SPSS verified the normal distribution of the data. Meanwhile, ANOVA was conducted on individual samples and the p values of ANOVA in testing procedures are consistent with that in the R simulation code.

Manual verification of the logic of the computer code was conducted. Individual samples were generated with small replications (20 trials) for both two-way and three-
way design. The number of rejections of each effect counted by the computer was compared to the results calculated by hand, which ensure the reasonable logic of the R coding. For example, the number of rejections in SFWER should be less than or equal to the rejections in TFWER. Or even if all the number of rejections in SFWER are all less than or equal to the rejections in TFWER, it doesn’t make sense that SFWER equals 0.0. As a result, the rejections from the computer were identical to the hand counting based on the \( p \) value matrix, in addition, no SFWER and TFWER equal 0. In addition, SFWER can also be verified based on known formulas. SFWER in R simulation code should be exactly the same as the formula of calculating the familywise Type I error rate when the null are completely true, \( p = 1 - (1 - \alpha)^k \), where \( \alpha \) is .05 and \( k \) is the number of independent hypothesis tests (Hochberg & Tamhane, 1987; Maxwell & Delaney, 2000; Schochet, 2008; Stevens, 1996; Toothaker, 1993).

Sensitivity testing ensures that the computer code is sensible with varied parameters (Brooks, 1998). For example, the three-way unbalanced design was tested. The descriptive statistics in the R testing code confirmed that the simulated data is normally distributed. The cell size in the simulation is consistent with the testing procedure.

The purpose of stress testing is to ensure that there are not unexpected results with strange values (Brooks, 1998). Varied numbers, for example, 0, 20, 100, and 1000 are used for the replications in the stress testing. As a result, the R simulation code runs well with these varied replications.
Data Analysis Procedure

The research questions of this study are: first, how these MHT procedures work in the context of factorial ANOVA (two-way and three-way) in terms of the Type I errors, second, how these MHT procedures work in the context of factorial ANOVA (two-way and three-way) in terms of statistical power rates. In order to answer these questions empirically, a Monte Carlo study is performed based on the design described in previous sections. Bradley (1978) provided two kinds of criteria to evaluate the robustness of departure from nominal $\alpha$: (a) the “fairly stringent criterion” defines the departure from $\alpha$ as $\alpha\pm1/10\alpha$, and (b) the “liberal criterion” defines the departure from $\alpha$ as $\alpha\pm1/2\alpha$. An “intermediate criterion” of $\alpha\pm1/4\alpha$, which was proposed by Brooks, Barcikowski, and Robey (1999), is also used in this study to evaluate the error rates. The following section describes how the Monte Carlo study answers these questions.

RQ1: How well do different MHT procedures control Type I errors in terms of Samplewise FWER, Testwise FWER, and FDR (a) in the balanced two-way factorial ANOVA and (b) in the balanced three-way factorial ANOVA?

Three measures were used to evaluate the Type I errors of the unadjusted alpha per test procedure and MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure). For the RQ1a, the results were presented as a 3x5 table. The rows are three patterns which include at least one true null hypothesis and the columns are five procedures. The Type I error rates were shown as Samplewise FWER, Testwise FWER, and FDR. A similar table was used for RQ1b which focused on the three-way design. The results were presented as a 7x5 table. The
rows are seven patterns which include at least one true null hypothesis and the columns are five procedures. It is necessary to note that there is no Type I error when all hypotheses are false null; therefore, three and seven patterns were used for the two-way and three-way analysis respectively.

RQ2: How well do different MHT procedures control statistical power in terms of Samplewise Power, Testwise Power, and TDR (a) in the balanced two-way factorial ANOVA and (b) in the balanced three-way factorial ANOVA?

These questions focused on the statistical power rates. Three measures were used to evaluate the statistical power of MHT procedures: Samplewise Power, Testwise Power, and TDR. For RQ2a, the results were presented as a 3x5 table. The rows are three patterns which include at least one false null hypothesis and the columns are five procedures. A 7x5 table was used for RQ2b which focuses on the three-way design. The rows are seven patterns which include at least one false null hypothesis and the columns are five procedures. It is necessary to note that there is no statistical power when all hypotheses are true null; therefore, three and seven patterns were used for the two-way and three-way designs respectively.

Chapter Summary

A Monte Carlo simulation was conducted in R program to investigate the Type I error rate and statistical power rate of several MHT procedures in the balanced two-way and three-way factorial ANOVA. Based on several patterns of true null and false null effects (four patterns in the two-way and eight patterns in the three-way), with medium effect size (.50) and fixed sample size per cell (32 per cell in the two-way and 16 in the
three-way), the Type I error rates (false rejections on true null hypotheses) were calculated in terms of Samplewise FWER, Testwise FWER, and FDR. Statistical power rates (true rejections on false null hypotheses) were calculated in terms of Samplewise Power, Testwise Power, and TDR, which have operational meaning in Monte Carlo simulation. Finally, each MHT procedure was evaluated based on each specific criterion.

The results were provided in the form of tables (a 3x5 table for the two-way design and a 7x5 table for the three-way design) in Chapter four, in which the performance of each procedure was reported. Readers can have a comprehensive and clear idea about the Type I error rates and statistical power rates of each MHT procedure in factorial designs. For example, what is the FDR in the Bonferroni Procedure? What is the FWER of the B-H procedure? The results part in Chapter four can answer these questions.
CHAPTER FOUR: RESULTS

The results from the Monte Carlo simulation were summarized in this chapter. Three indicators of Type I error rates (Samplewise FWER, Testwise FWER, and FDR) and three criteria of statistical power rates (Samplewise Power, Testwise Power, and TDR) of the unadjusted alpha per test procedure and four MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure) were displayed in tables and figures. The sequences of the results were based on the two-way and three-way design respectively.

Type I Error Rate in the Two-way Balanced Design

The research question (RQ1a) was: How well do different MHT procedures control Type I error rates in terms of Samplewise FWER (SFWER), Testwise FWER (TFWER), and FDR in the balanced two-way factorial ANOVA? The results from 20,000 Monte Carlo replications in the two-way balanced design are displayed in Table 9. There are four patterns of true null and false null effects in the two-way balanced design. Type I error rates are calculated for three patterns that include at least one true null hypothesis.
Table 4

*Type I Error Rates in Two-Way Balanced Design with 20,000 Replications*

<table>
<thead>
<tr>
<th>Number of True Null</th>
<th>Error Rates</th>
<th>Unadjusted</th>
<th>Bonferroni</th>
<th>Holm</th>
<th>Hochberg</th>
<th>B-H</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>SFWER</td>
<td>0.1393</td>
<td>0.0476</td>
<td>0.0476</td>
<td>0.0478</td>
<td>0.0483</td>
</tr>
<tr>
<td></td>
<td>TFWER</td>
<td>0.0486</td>
<td>0.0162</td>
<td>0.0164</td>
<td>0.0166</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>SFWER</td>
<td>0.0974</td>
<td>0.0329</td>
<td>0.0440</td>
<td>0.0456</td>
<td>0.0578</td>
</tr>
<tr>
<td></td>
<td>TFWER</td>
<td>0.0499</td>
<td>0.0166</td>
<td>0.0228</td>
<td>0.0239</td>
<td>0.0301</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.1106</td>
<td>0.0479</td>
<td>0.0645</td>
<td>0.0674</td>
<td>0.0829</td>
</tr>
<tr>
<td>1</td>
<td>SFWER</td>
<td>0.0509</td>
<td>0.0166</td>
<td>0.0371</td>
<td>0.0416</td>
<td>0.0438</td>
</tr>
<tr>
<td></td>
<td>TFWER</td>
<td>0.0509</td>
<td>0.0166</td>
<td>0.0371</td>
<td>0.0416</td>
<td>0.0438</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.0307</td>
<td>0.0124</td>
<td>0.0260</td>
<td>0.0288</td>
<td>0.0291</td>
</tr>
</tbody>
</table>

*Three True Null Hypotheses*

As introduced in Chapter one, SFWER is the probability of at least one false rejection on true null hypotheses in terms of samples that are generated in the simulation. Under the pattern of three completely true null hypotheses, the SFWER in the unadjusted alpha procedure is .1393, which is consistent with the formula of $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .05 and $k$ is 3 ($k$ is the number of independent hypothesis test). Meanwhile, it is far beyond the liberal criterion interval, $[.025, .075]$, $\alpha \pm 1/2 \alpha$, (Bradley, 1978). The SFWER result from MHT procedures are within the interval of $[.045, .055]$ based on the stringent criterion, $\alpha \pm 1/10 \alpha$, (Bradley, 1978). The SFWER in the Bonferroni procedure is also consistent with the formula of $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .0167 (.05/3) and $k$ is 3.

TFWER is the probability of at least one false rejection on true null hypotheses in terms of all tests in the simulation. The TFWER of the unadjusted alpha procedure is about .0486. In MHT procedures, the TFWER are almost the same. FDR is the expected
proportion of the false rejections and total rejections. With three true null hypotheses, FDR is 1.0 in all procedures, which is consistent with the property of FDR that was illustrated in Chapter one: when the number of false rejections is the same as the total rejections, FDR is equal to 1.0.

Two True Null Hypotheses

When there are two true null hypotheses and one false null, the SFWER in the unadjusted procedure is about .0974, which is also consistent with the formula of

\[ p = 1 - (1 - \alpha)^k \] where \( \alpha \) is .05 and \( k \) is 2. Meanwhile, it is beyond the liberal criterion interval, \([.025, .075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978). In MHT procedures, the Bonferroni procedure has the SFWER of .0329, which is consistent with the formula of

\[ p = 1 - (1 - \alpha)^k \] where \( \alpha \) is .0167 and \( k \) is 2. Meanwhile, .0329 is in the interval of \([.025, .075]\) based on the liberal criterion, \(\alpha \pm 1/2\alpha\), (Bradley, 1978). The SFWER in the other three MHT procedures are within the interval of the intermediate criterion, \([.0375, .0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). The TFWER in the unadjusted procedure is .0499. In MHT procedures, the Bonferroni procedure has the TFWER of .0166 and the B-H procedure has the TFWER of .0301. The TFWER in the Holm approach and the Hochberg procedure is .0228 and .0239, respectively. The FDR in the unadjusted procedure is .1106. In MHT procedures, the FDR in the Bonferroni procedure is about .0479. The B-H procedure has the FDR of .0829. The FDR in the Hochberg procedure is slightly greater than the Holm procedure, which is .0674 and .0645, respectively.
One True Null Hypothesis

When there is only one true null hypothesis, the SFWER in the unadjusted procedure is about .0509, which is consistent with the formula of \( p = 1 - (1 - \alpha)^k \) where \( \alpha \) is .05 and \( k \) is 1. Meanwhile, .0509 is within the interval of [.045, .055] based on the stringent criterion, \( \alpha \pm 1/10\alpha \), (Bradley, 1978). The SFWER in the Bonferroni procedure is about .0166, which is consistent with the formula of \( p = 1 - (1 - \alpha)^k \) where \( \alpha \) is .05/3 and \( k \) is 1. Meanwhile, .0166 is not in the interval of [.025, .075] based on the liberal criterion, \( \alpha \pm 1/2\alpha \), (Bradley, 1978). The SFWER in the Holm procedure, the Hochberg procedure, and B-H approach are within the interval of the intermediate criterion, [.0375, .0625], \( \alpha \pm 1/4\alpha \), (Brooks, Barcikowski, & Robey, 1999). The TFWER in each procedure is the same as the SFWER because there is only one true null hypothesis. The FDR in the unadjusted approach is about .0307. The Bonferroni procedure has FDR of .0124 and .0260 is in the Holm procedure. The FDR is .0288 in the Hochberg procedure and .0291 is in the B-H procedure. Figure 1 to Figure 3 show the Type I error rates in each procedure in the two-way design with 20,000 replications.
It is obvious to see that the SFWER in the unadjusted procedure increases greatly with multiple true null hypotheses. The SFWER in the Bonferroni procedure increases with more true null hypotheses, from .0166 to .0476. The SFWER in the Holm procedure and the Hochberg procedure seem stable in each pattern, the gap between the two procedures is very small. When there are two true null hypotheses, the SFWER in the B-H procedure is the highest, .0578. With one true and three true null hypotheses, the SFWER are almost the same. Graphically, it is symmetrical.
Figure 2. Testwise FWER in two-way balanced design with 20,000 replications.

The TFWER in the unadjusted procedure is constant in each pattern, which is .05. The line of the unadjusted procedure locates higher than MHT procedures. In the Bonferroni procedure, the TFWER also looks stable in each pattern, which is about .0166. Meanwhile, the Bonferroni line is the lowest in the graph. With more true null hypotheses, the TFWER rates in the Holm procedure, the Hochberg procedure, and the B-H procedure decrease. The B-H line is higher than the Hochberg procedure; the Hochberg line is higher than the Holm procedure. The TFWER rates are almost the same with three true null hypotheses in MHT procedures.
Figure 3. FDR in two-way balanced design with 20,000 replications.

The FDR increases in all procedures with more true null hypotheses. The unadjusted procedure has higher FDR than MHT procedures and accordingly, the line of the unadjusted procedure is higher than MHT procedures. The B-H procedure, the Hochberg procedure, the Holm procedure, and the Bonferroni procedure locate from high to low in order.

Statistical Power Rate in the Two-way Balanced Design

The research question (RQ2a) was: How well do different MHT procedures show statistical power rates in terms of Samplewise Power, Testwise Power, and TDR in the balanced two-way factorial ANOVA? There are four patterns of true null and false null effects in the two-way balanced design. Statistical power rates are calculated for three patterns which include at least one false null hypothesis. Table 10 summarizes the statistical power rates with 20,000 replications, the Samplewise Power (SPOWER), Testwise Power (TPOWER), and TDR are calculated for each procedure.
Table 5

Statistical Power Rates in Two-Way Balanced Design with 20,000 Replications

<table>
<thead>
<tr>
<th>Number of False Null</th>
<th>Power Rates</th>
<th>Unadjusted</th>
<th>Bonferroni</th>
<th>Holm</th>
<th>Hochberg</th>
<th>B-H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SPOWER</td>
<td>0.8029</td>
<td>0.6578</td>
<td>0.6600</td>
<td>0.6612</td>
<td>0.6646</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8029</td>
<td>0.6578</td>
<td>0.6600</td>
<td>0.6612</td>
<td>0.6646</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.8894</td>
<td>0.9521</td>
<td>0.9355</td>
<td>0.9326</td>
<td>0.9171</td>
</tr>
<tr>
<td>2</td>
<td>SPOWER</td>
<td>0.9572</td>
<td>0.8784</td>
<td>0.8790</td>
<td>0.8828</td>
<td>0.8878</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8029</td>
<td>0.6586</td>
<td>0.6956</td>
<td>0.7011</td>
<td>0.7297</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.9693</td>
<td>0.9876</td>
<td>0.9740</td>
<td>0.9712</td>
<td>0.9709</td>
</tr>
<tr>
<td>3</td>
<td>SPOWER</td>
<td>0.9919</td>
<td>0.9578</td>
<td>0.9578</td>
<td>0.9629</td>
<td>0.9659</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.7997</td>
<td>0.6535</td>
<td>0.7474</td>
<td>0.7622</td>
<td>0.7745</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

One False Null Hypothesis

SPOWER is the probability of at least one true rejection on false null hypotheses in terms of samples that are generated in the simulation. When there is only one false null hypothesis and two true null hypotheses, the unadjusted procedure has the SPOWER of .8029. The Bonferroni procedure has the smallest SPOWER of .6578 and the B-H procedure has the highest of .6646 in MHT procedures. The SPOWER in the Holm procedure is .6600 and the Hochberg procedure has a slightly higher value of .6612. The TPOWER is the probability of detecting all true rejections on false null hypotheses in terms of all tests in the simulation. The TPOWER is the same as SPOWER in each procedure because there is only one false null hypothesis. The TDR is the expected proportion of the number of true rejections and the total rejections. The TDR in the unadjusted procedure is .8894. The Bonferroni procedure has the TDR of .9521 while the
B-H procedure has the TDR of .9171. The TDR in the Holm procedure and Hochberg procedures is .9355 and .9326, respectively.

*Two False Null Hypotheses*

When there are two false null hypotheses and one true null hypothesis, the SPOWER in the unadjusted procedure is about .9572. In the four MHT procedures, the Bonferroni procedure has the smallest SPOWER of .8784 and the B-H approach has the highest SPOWER of .8878. The SPOWER in the Hochberg procedure is slightly higher than the Holm procedure. The TPOWER in the unadjusted procedure is .8029. In MHT procedures, the pattern of TPOWER is similar to SPOWER. That is, the Bonferroni procedure has the smallest TPOWER and the B-H procedure has the highest TPOWER, while the Hochberg procedure and the Holm procedure rank the second and the third respectively. The TDR in the B-H procedure is the smallest and in the Bonferroni procedure is the largest, the Holm procedure and the Hochberg procedure rank the second and the third respectively.

*Three False Null Hypotheses*

When there are three false null hypotheses, the SPOWER and TPOWER in the unadjusted procedure is .9919 and .7997 respectively. The SPOWER in the Bonferroni procedure is the same as that in the Holm procedure, which is .9578. The TPOWER in the Bonferroni procedure is smaller than that in the Holm procedure. The B-H procedure has the largest SPOWER and TPOWER in MHT procedures. The SPOWER and TPOWER in the Hochberg procedure are slightly higher than the Holm procedure. The
TDR in all procedures is 1.0 because all hypotheses are false in this pattern. The following Figures show the statistical power rates in the two-way design.

*Figure 4.* Samplewise Power in two-way balanced design with 20,000 replications.

With more false null hypotheses, the SPOWER increases in each procedure. The unadjusted procedure has greater SPOWER than MHT procedures and accordingly, the line of unadjusted procedure locates upper than MHT procedures. In each pattern, the SPOWER of MHT procedures are almost the same; the B-H procedure seems to have a little bit more power advantage than other MHT procedures in graph. Overall, the lines indicating MHT procedures are clustered together and almost coincide with each other.
Figure 5. Testwise Power in two-way balanced design with 20,000 replications.

The TPOWER rates seem stable in the unadjusted procedure and the Bonferroni procedure, which locate the highest and lowest in the above figure. With more false null hypotheses, the TPOWER rates increase in the Holm procedure, the Hochberg procedure, and the B-H procedure. Specifically, the TPOWER of the B-H procedure is the highest, the Hochberg procedure and the Holm procedure rank the second and the third respectively.
Figure 6. TDR in two-way balanced design with 20,000 replications.

With more false null hypotheses, TDR in each procedure increases. The TDR is 1.0 with three false null hypotheses. The TDR in the Bonferroni procedure is the highest and accordingly, the Bonferroni line locates the highest in the above figure. The Holm procedure and the Hochberg procedure seem coinciding to each other. The B-H procedure has the lowest TDR in MHT procedures. The unadjusted procedure has lower TDR than MHT procedures. Meanwhile, the differences of TDR among procedures become very small with the number of false null hypotheses increases.

Type I Error Rate in the Three-way Balanced Design

The research question (RQ1b) was: How well do different MHT procedures control Type I error rates in terms of Samplewise FWER (SFWER), Testwise FWER (TFWER), and FDR in the balanced three-way factorial ANOVA? The results from 20,000 Monte Carlo replications in the three-way balanced design are displayed in Table 11.
Table 6

<table>
<thead>
<tr>
<th>Number of True Null</th>
<th>Error Rates</th>
<th>Unadjusted</th>
<th>Bonferroni</th>
<th>Holm</th>
<th>Hochberg</th>
<th>B-H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SFWER</td>
<td>0.2913</td>
<td>0.0479</td>
<td>0.0479</td>
<td>0.0479</td>
<td>0.0491</td>
</tr>
<tr>
<td>7</td>
<td>TFWER</td>
<td>0.0492</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0071</td>
<td>0.0078</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.2627</td>
<td>0.0400</td>
<td>0.0443</td>
<td>0.0446</td>
<td>0.0682</td>
</tr>
<tr>
<td>6</td>
<td>TFWER</td>
<td>0.0504</td>
<td>0.0068</td>
<td>0.0077</td>
<td>0.0077</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.2731</td>
<td>0.0702</td>
<td>0.0782</td>
<td>0.0787</td>
<td>0.1203</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.2217</td>
<td>0.0360</td>
<td>0.0427</td>
<td>0.0431</td>
<td>0.0809</td>
</tr>
<tr>
<td>5</td>
<td>TFWER</td>
<td>0.0493</td>
<td>0.0073</td>
<td>0.0088</td>
<td>0.0089</td>
<td>0.0175</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.1334</td>
<td>0.0326</td>
<td>0.0384</td>
<td>0.0387</td>
<td>0.0674</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.1883</td>
<td>0.0297</td>
<td>0.0404</td>
<td>0.0411</td>
<td>0.0909</td>
</tr>
<tr>
<td>4</td>
<td>TFWER</td>
<td>0.0509</td>
<td>0.0075</td>
<td>0.0103</td>
<td>0.0105</td>
<td>0.0244</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.0780</td>
<td>0.0182</td>
<td>0.0237</td>
<td>0.0240</td>
<td>0.0473</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.1396</td>
<td>0.0201</td>
<td>0.0334</td>
<td>0.0346</td>
<td>0.0803</td>
</tr>
<tr>
<td>3</td>
<td>TFWER</td>
<td>0.0492</td>
<td>0.0068</td>
<td>0.0115</td>
<td>0.0120</td>
<td>0.0283</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.0442</td>
<td>0.0093</td>
<td>0.0147</td>
<td>0.0152</td>
<td>0.0298</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.0934</td>
<td>0.0140</td>
<td>0.0277</td>
<td>0.0298</td>
<td>0.0645</td>
</tr>
<tr>
<td>2</td>
<td>TFWER</td>
<td>0.0482</td>
<td>0.0070</td>
<td>0.0143</td>
<td>0.0158</td>
<td>0.0335</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.0234</td>
<td>0.0051</td>
<td>0.0094</td>
<td>0.0103</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td>SFWER</td>
<td>0.0511</td>
<td>0.0065</td>
<td>0.0208</td>
<td>0.0270</td>
<td>0.0407</td>
</tr>
<tr>
<td>1</td>
<td>TFWER</td>
<td>0.0511</td>
<td>0.0065</td>
<td>0.0208</td>
<td>0.0270</td>
<td>0.0407</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>0.0105</td>
<td>0.0020</td>
<td>0.0055</td>
<td>0.0070</td>
<td>0.0090</td>
</tr>
</tbody>
</table>

Seven True Null Hypotheses

Under the pattern of seven completely true null hypotheses, the SFWER in the unadjusted alpha per test procedure is about .2913, which is consistent with the formula of $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .05 and $k$ is 7. Meanwhile, .2913 is far beyond the liberal criterion interval of [.025, .075], $\alpha \pm 1/2\alpha$, (Bradley, 1978). The SFWER in MHT
procedures are within the interval of stringent criterion, \([.045, .055], \alpha \pm 1/10\alpha\), (Bradley, 1978). Specifically, the SFWER in the Bonferroni procedure is consistent with the formula of \( p = 1 - (1 - \alpha)^k \) where \( \alpha \) is .0071 (.05/7) and \( k \) is 7. The unadjusted procedure has the TFWER of .0492. The TFWER rates result from MHT procedures are almost the same. The FDR in all procedures is 1.0 because all hypotheses in this condition are true null hypotheses.

**Six True Null Hypotheses**

Under the pattern of six true null hypotheses and one false null hypothesis, the SFWER in the unadjusted procedure is about .2627, which is consistent with the formula of \( p = 1 - (1 - \alpha)^k \) where \( \alpha \) is .05 and \( k \) is 6. Meanwhile, .2627 is far beyond the liberal criterion interval of \([.025, .075], \alpha \pm 1/2\alpha\), (Bradley, 1978). The SFWER in the Bonferroni procedure, the Holm procedure, and the Hochberg procedure are within the interval of \([.0375, .0625], \alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). The SFWER in the B-H procedure is in the liberal interval of \([.025, .075], \alpha \pm 1/2\alpha\), (Bradley, 1978). The TFWER in the unadjusted procedure is about .0504. The Bonferroni procedure has the TFWER of .0068 and the B-H procedure has the TFWER of .0125. The TFWER in the Hochberg is the same as the Holm procedure, which is .0077. The FDR of the unadjusted procedure is about .2731. The Bonferroni procedure has the smallest FDR and the B-H procedure is the largest in MHT procedures. The Hochberg has a little bit higher FDR than the Holm procedure.

**Five True Null Hypotheses**
When there are five true null hypotheses and two false null hypotheses, the SFWER in the unadjusted procedure is about .2217, which is identical to the formula of 

\[ p = 1 - (1 - \alpha)^k \]

where \( \alpha \) is .05 and \( k \) is 5. Meanwhile, .2217 is far beyond the liberal criterion interval of [.025, .075], \( \alpha \pm 1/2 \alpha \), (Bradley, 1978). The SFWER results from the Bonferroni procedure is .036, which is in the liberal interval of [.025, .075], \( \alpha \pm 1/2 \alpha \), (Bradley, 1978). The SFWER in the Holm procedure and the Hochberg procedure are within the intermediate interval of [.0375, .0625], \( \alpha \pm 1/4 \alpha \), (Brooks, Barcikowski, & Robey, 1999). The B-H procedure has the SFWER of .0809 which is beyond the liberal criterion interval of [.025, .075], \( \alpha \pm 1/2 \alpha \), (Bradley, 1978). The TFWER in the unadjusted procedure is about .0493. The Bonferroni procedure has the TFWER of .0073 and the B-H approach has .0175. The TFWER in the Holm procedure and the Hochberg procedure is .0088 and .0089 respectively. The FDR in the unadjusted procedure is .1334. The Bonferroni procedure has the FDR of .0326 and .0384 is in the Holm procedure. The Hochberg has slightly higher FDR than the Holm procedure. The B-H procedure has the highest FDR of .0674.

**Four True Null Hypotheses**

Under the pattern of four true null hypotheses and three false null, the SFWER in the unadjusted is about .1883 which is close to the formula of 

\[ p = 1 - (1 - \alpha)^k \]

where \( \alpha \) is .05 and \( k \) is 4. In fact, .1883 is far beyond the liberal criterion interval of [.025, .075], \( \alpha \pm 1/2 \alpha \), (Bradley, 1978). The SFWER in the Bonferroni procedure is .0297, which falls in the liberal interval of [.025, .075], \( \alpha \pm 1/2 \alpha \), (Bradley, 1978). The SFWER in the Holm procedure and the Hochberg procedure are in the intermediate interval of [.0375, .0625],
$\alpha \pm 1/4\alpha$, (Brooks, Barcikowski, & Robey, 1999). The B-H procedure has the SFWER of .0909 which is beyond the liberal criterion interval of [.025, .075], $\alpha \pm 1/2\alpha$, (Bradley, 1978). The TFWER in the unadjusted procedure is about .0509. The Bonferroni procedure has the TFWER of .0075 and the B-H has .0244. The TFWER in the Holm procedure and the Hochberg procedure is .0103 and .0105 respectively. The FDR in the unadjusted procedure is about .078. The Bonferroni procedure has the smallest value of FDR and the B-H procedure has the largest FDR in MHT procedures. The Hochberg procedure has slightly greater FDR than the Holm procedure.

**Three True Null Hypotheses**

When there are three true null hypotheses and four false null, The SFWER in the unadjusted procedure is .1396, which is identical to the formula of $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .05 and $k$ is 3. In fact, .1396 is beyond the liberal criterion interval of [.025, .075], $\alpha \pm 1/2\alpha$, (Bradley, 1978). The SFWER in the Bonferroni procedure and the B-H procedure are beyond the liberal criterion interval of [.025, .075], $\alpha \pm 1/2\alpha$, (Bradley, 1978). The SFWER in the Holm procedure and the Hochberg procedure are within the liberal interval of [.025, .075], $\alpha \pm 1/2\alpha$, (Bradley, 1978). The TFWER in the unadjusted procedure is .0492. The Bonferroni procedure has the TFWER of .0068, .0115 is in the Holm procedure, .012 in the Hochberg procedure, and .0283 in the B-H procedure. The FDR in the unadjusted procedure is .0442. The Bonferroni procedure has the smallest FDR and the B-H procedure has the largest. The Hochberg procedure and the Holm procedure rank the second and the third.
Two True Null Hypotheses

When there are two true null hypotheses and five false null, The SFWER in the unadjusted procedure is .0934, which is identical to the formula of $p = 1 - (1 - \alpha)^k$ where $\alpha$ is .05 and $k$ is 2. In fact, .0934 is beyond the liberal criterion interval of [.025, .075], $\alpha \pm \frac{1}{2}\alpha$, (Bradley, 1978). The SFWER in the Bonferroni procedure is beyond the liberal criterion interval of [.025, .075], $\alpha \pm \frac{1}{2}\alpha$, (Bradley, 1978). The SFWER in the Holm procedure, the Hochberg procedure, and the B-H procedure are within the liberal interval of [.025, .075], $\alpha \pm \frac{1}{2}\alpha$, (Bradley, 1978). The TFWER in the unadjusted procedure is .0482. The Bonferroni procedure has the TFWER of .007, .0143 is in the Holm procedure, .0158 in the Hochberg procedure, and .0335 in the B-H procedure. The FDR in the unadjusted procedure is .0234. The Bonferroni procedure has the smallest FDR and the B-H has the largest. The Hochberg procedure and the Holm procedure rank the second and the third.

One True Null Hypothesis

When there is only one true null hypothesis, the SFWER is the same as TFWER in each procedure. The unadjusted procedure has the SFWER of .0511 which is in the stringent criterion, [.045, .055], $\alpha \pm \frac{1}{10}\alpha$, (Bradley, 1978). The SFWER in the Bonferroni procedure and the Holm procedure are out of the liberal interval of [.025, .075], $\alpha \pm \frac{1}{2}\alpha$, (Bradley, 1978). The SFWER in the Hochberg procedure is in the liberal interval of [.025, .075], $\alpha \pm \frac{1}{2}\alpha$, (Bradley, 1978) and the B-H procedure is in the intermediate interval of [.0375, .0625], $\alpha \pm \frac{1}{4}\alpha$, (Brooks, Barcikowski, & Robey, 1999). The FDR in the unadjusted procedure is .015. The Bonferroni procedure has the smallest FDR while
the B-H procedure has the largest in MHT procedures. The FDR in the Hochberg procedure is slightly higher than the Holm procedure. The following Figures display the Type I error rates of all procedures in the three-way design.

![Samplewise FWER in three-way balanced design with 20,000 replications.](image)

*Figure 7. Samplewise FWER in three-way balanced design with 20,000 replications.*

The line indicating the unadjusted procedure locates the highest. It is obvious to see that the SFWER in the unadjusted procedure increase greatly with the number of true null hypotheses increases. The SFWER in the Bonferroni procedure increase with the number of true null increases, from .0065 to .0479. The line of the Bonferroni procedure locates the lowest. The SFWER in the Holm procedure and the Hochberg procedure seem stable and do not change very much with more true null hypotheses. In fact, the two lines are almost coinciding. The SFWER in the B-H procedure reaches the highest with four true null hypotheses, from one true null to four true null hypotheses, the SFWER
increase; from four true null to seven true null, the SFWER decrease. The distribution of SFWER in the B-H procedure is graphically symmetrical.

![Graph showing testwise FWER in three-way balanced design with 20,000 replications.](image)

**Figure 8.** Testwise FWER in three-way balanced design with 20,000 replications.

The line showing the unadjusted procedure locates higher than MHT procedures. The Bonferroni procedure line is the lowest in the graph. The TFWER rates in the unadjusted procedure and the Bonferroni procedure seem stable, which are .05 and .0166 respectively. With more true null hypotheses, the TFWER rates in the Holm procedure, the Hochberg procedure, and the B-H procedure decrease. The B-H procedure line is upper than other MHT procedures; the Holm procedure and the Hochberg procedure are almost coinciding, except that there are small differences from three true null hypotheses to one true null hypothesis.
Figure 9. FDR in three-way balanced design with 20,000 replications.

The FDR in each procedure decreases with more false null hypotheses. The line of unadjusted procedure is above MHT procedures. The B-H line is higher than other MHT procedures. The Holm procedure and the Hochberg procedure are almost coinciding. The Bonferroni line is the lowest in graph.

Statistical Power Rate in the Three-way Balanced Design

The research question (RQ2b) was: How well do different MHT procedures show statistical power rate in terms of Samplewise Power, Testwise Power, and TDR in the balanced three-way factorial ANOVA? There are eight patterns of true null and false null effects in the three-way balanced design. Statistical power rate is calculated for seven patterns which include at least one false null hypothesis. Table 12 summarizes the statistical power rates with 20,000 replications, the Samplewise Power (SPOWER), Testwise Power (TPOWER), and TDR are calculated in each procedure.
### Table 7

**Statistical Power Rates in Three-Way Balanced Design with 20,000 Replications**

<table>
<thead>
<tr>
<th>Number of False Null</th>
<th>Power Rates</th>
<th>Unadjusted</th>
<th>Bonferroni</th>
<th>Holm</th>
<th>Hochberg</th>
<th>B-H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPOWER</td>
<td>0.8041</td>
<td>0.5407</td>
<td>0.5417</td>
<td>0.5419</td>
<td>0.5493</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8041</td>
<td>0.5407</td>
<td>0.5417</td>
<td>0.5419</td>
<td>0.5493</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.7269</td>
<td>0.9298</td>
<td>0.9218</td>
<td>0.9213</td>
<td>0.8797</td>
</tr>
<tr>
<td>2</td>
<td>SPOWER</td>
<td>0.9607</td>
<td>0.7842</td>
<td>0.7849</td>
<td>0.7856</td>
<td>0.7987</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8005</td>
<td>0.5411</td>
<td>0.5523</td>
<td>0.5529</td>
<td>0.6061</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.8666</td>
<td>0.9674</td>
<td>0.9616</td>
<td>0.9613</td>
<td>0.9326</td>
</tr>
<tr>
<td>3</td>
<td>SPOWER</td>
<td>0.9922</td>
<td>0.8949</td>
<td>0.8952</td>
<td>0.8961</td>
<td>0.9105</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8022</td>
<td>0.5408</td>
<td>0.5669</td>
<td>0.5684</td>
<td>0.6540</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.9220</td>
<td>0.9818</td>
<td>0.9763</td>
<td>0.9760</td>
<td>0.9527</td>
</tr>
<tr>
<td>4</td>
<td>SPOWER</td>
<td>0.9979</td>
<td>0.9469</td>
<td>0.9470</td>
<td>0.9475</td>
<td>0.9596</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.7980</td>
<td>0.5373</td>
<td>0.5786</td>
<td>0.5814</td>
<td>0.6893</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.9558</td>
<td>0.9907</td>
<td>0.9853</td>
<td>0.9848</td>
<td>0.9702</td>
</tr>
<tr>
<td>5</td>
<td>SPOWER</td>
<td>0.9997</td>
<td>0.9761</td>
<td>0.9761</td>
<td>0.9766</td>
<td>0.9839</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8035</td>
<td>0.5411</td>
<td>0.6011</td>
<td>0.6063</td>
<td>0.7247</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.9766</td>
<td>0.9949</td>
<td>0.9906</td>
<td>0.9897</td>
<td>0.9818</td>
</tr>
<tr>
<td>6</td>
<td>SPOWER</td>
<td>0.9999</td>
<td>0.9859</td>
<td>0.9859</td>
<td>0.9865</td>
<td>0.9924</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8025</td>
<td>0.5395</td>
<td>0.6239</td>
<td>0.6361</td>
<td>0.7495</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>0.9895</td>
<td>0.9980</td>
<td>0.9945</td>
<td>0.9930</td>
<td>0.9910</td>
</tr>
<tr>
<td>7</td>
<td>SPOWER</td>
<td>1.0000</td>
<td>0.9940</td>
<td>0.9940</td>
<td>0.9947</td>
<td>0.9969</td>
</tr>
<tr>
<td></td>
<td>TPOWER</td>
<td>0.8033</td>
<td>0.5408</td>
<td>0.6605</td>
<td>0.6894</td>
<td>0.7729</td>
</tr>
<tr>
<td></td>
<td>TDR</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The rates of SPOWER and TPOWER in the unadjusted procedure under all patterns are adequate enough to or greater than .80. The Bonferroni procedure has the smallest SPOWER and TPOWER while the B-H procedure has the greatest in MHT procedures. The SPOWER and TPOWER in the Hochberg procedure are slightly greater than the Holm procedure. The TDR in the Bonferroni procedure is the greatest while the B-H procedure is the smallest. Meanwhile, the TDR in all procedures is 1.0 under the
pattern of seven false null hypotheses. In addition, when there are six true null hypotheses and one false null, the SPOWER and TPOWER is the same in each procedure.

For example, when there are five true null hypotheses and two false null, the SPOWER in the unadjusted procedure is .9607. In MHT procedures, the Bonferroni procedure has the smallest SPOWER of .7842 while the B-H procedure has the highest .7987. The SPOWER in the Holm procedure and in the Hochberg procedure is .7849 and .7856 respectively. The TPOWER in the unadjusted procedure is .8005. In MHT procedures, the TPOWER in the Bonferroni procedure is the smallest and the B-H procedure is the greatest. The TDR in the Bonferroni procedure is .9674 and .9326 in the B-H procedure. The following Figures show the statistical power rates of all procedures in the three-way design.

Figure 10. Samplewise Power in three-way balanced design with 20,000 replications.
The line indicating the unadjusted procedure is above MHT procedures. The SPOWER is very close to 1.0 with seven false null hypotheses in all procedure. The B-H line locates higher than other MHT procedures, which indicates more power. The SFWER in other three MHT procedures are almost the same and the three lines are coinciding.

![Graph showing testwise power](image)

**Figure 11.** Testwise Power in three-way balanced design with 20,000 replications.

The TPOWER rates seem stable in the unadjusted procedure and the Bonferroni procedure, which are the highest and lowest in the above figure. With more false null hypotheses, the TPOWER rates increase in the Holm procedure, the Hochberg procedure, and the B-H procedure. Specifically, the TPOWER of the B-H procedure is the highest, the Hochberg procedure and the Holm procedure rank the second and the third respectively. Meanwhile, the TPOWER difference between the Holm procedure and the Hochberg procedure looks very small.
Figure 12. TDR in three-way balanced design with 20,000 replications.

With more false null hypotheses, TDR in each procedure increases. The TDR is 1.0 with seven false null hypotheses. The TDR in the Bonferroni procedure is the highest in the above figure and accordingly, the Bonferroni line is the highest. The Holm procedure and the Hochberg procedure are the second and the third, in fact, the two lines are coinciding. The B-H procedure has the lowest TDR in MHT procedures. The unadjusted procedure has lower TDR than MHT procedures and consequently, the unadjusted line is the lowest in the graph. Meanwhile, the differences of TDR among procedures become very small with more false null hypotheses.

Summary of Type I Error Rates

Performance of the Unadjusted Alpha Per Test Procedure

The SFWER in the unadjusted alpha per test procedure is around .05 with a single hypothesis test, which is within the stringent interval of [.045, .055], $\alpha \pm 1/10\alpha$, (Bradley,
1978). For example, the SFWER in the unadjusted procedure is .0509 when there is only one true null hypothesis in the two-way design and .0511 for the three-way design. The SFWER in the unadjusted procedure increase greatly with multiple hypothesis tests. For example, the SFWER is .0974 with two true null hypotheses and .1393 with three hypotheses in the two-way design. The SFWER becomes .2913 with seven hypotheses in the three-way design. The inflated SFWER are beyond the liberal interval of \([.025, .075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978). The TFWER in the unadjusted procedure is around .05 in each pattern of true null and false null effect in both two-way and three-way designs. The FDR in the unadjusted procedure is greater than MHT procedures. Meanwhile, the FDR decreases as the number of true null hypotheses decreases.

**Performance of the Bonferroni Procedure**

Several SFWER result from the Bonferroni procedure are within the liberal interval of \([.025, .075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978). Specifically, some are even in the intermediate interval of \([.0375, .0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999), or in the stringent interval of \([.045, .055]\), \(\alpha \pm 1/10\alpha\), (Bradley, 1978). For example, the SFWER is .0479 with seven true null hypotheses in the three-way design, which is in the stringent interval (Bradley, 1978). The SFWER of .04 with six true null hypotheses in the three-way design is in the intermediate interval (Brooks, Barcikowski, & Robey, 1999). The SFWER of .0202, .014, and .0065 in the three-way design and .0166 in the two-way design are beyond the liberal interval of \([.025, .075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978). The TFWER rates result from the Bonferroni procedure are very stable in each pattern in both two-way and three-way designs, .0166 and .0071 respectively. The FDR rates in the
Bonferroni procedure are generally the smallest in MHT procedures. Meanwhile, the FDR decreases as the number of true null hypotheses decreases.

**Performance of the Holm Procedure**

All SFWER in the two-way design of the Holm procedure are within the intermediate interval of \([0.0375, 0.0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). In the three-way design, six SFWER (0.0277, 0.0334, 0.0404, 0.0427, 0.0443, and 0.0479) are within liberal interval of \([0.025, 0.075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978) except that 0.0208 is out of this interval. Moreover, 0.0404, 0.0427, 0.0443, and 0.0479 are in the intermediate interval of \([0.0375, 0.0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). The TFWER rates in the Holm procedure increase a little bit with less true null hypotheses. The FDR rates are greater than the Bonferroni procedure. Meanwhile, the FDR decreases as the number of true null hypotheses decreases.

**Performance of the Hochberg Procedure**

All SFWER result from the Hochberg procedure are within liberal interval of \([0.025, 0.075]\), \(\alpha \pm 1/2\alpha\), (Bradley, 1978). Some are even in the stringent interval of \([0.045, 0.055]\), \(\alpha \pm 1/10\alpha\), (Bradely, 1978); some are in the intermediate interval of \([0.0375, 0.0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). For example, in the three-way design, 0.027, 0.0298, and 0.0346 are in the liberal interval; 0.0411 and 0.0431 are in the intermediate criterion interval; 0.0446 and 0.0479 are in the stringent interval. All SFWER in the two-way design are in the intermediate interval of \([0.0375, 0.0625]\), \(\alpha \pm 1/4\alpha\), (Brooks, Barcikowski, & Robey, 1999). The TFWER rates increase with the number of true null hypotheses decreases. The FDR in this procedure is a little bit greater than the Bonferroni
procedure and the Holm procedure, but less than the B-H procedure. Meanwhile, the FDR decreases as the number of true null hypotheses decreases.

*Performance of the B-H Procedure*

Some SFWER result from the B-H procedure are within the liberal interval of $[.025, .075]$, $\alpha \pm 1/2\alpha$, (Bradley, 1978), some are not. For example, all SFWER in the two-way design and .0407, .0645, .0682, and .0491 in the three-way design are in the liberal interval; .0809, .0909, and .0803 are beyond the liberal interval of $[.025, .075]$, $\alpha \pm 1/2\alpha$, (Bradley, 1978). In the two-way design, the SFWER is the greatest with two true null hypotheses; for one true null and three true null hypotheses, the SFWER are almost the same. In the three-way design, the SFWER is the greatest with four true null hypotheses. The SFWER are almost the same for one true and seven true null hypotheses, two true and six true null hypotheses; and three true and five true null hypotheses. The TFWER rates increase with the number of true null hypotheses decreases. The FDR in the B-H procedure are the greatest in MHT procedures. Meanwhile, the FDR decreases as the number of true null hypotheses decreases. For example, the FDR in the B-H procedure is .1203 with one false null and six true null hypotheses in the three-way design, it decreases to .0674 with two false null and five true null hypotheses. The FDR is .0473 with three false null and four true null hypotheses.

*Summary of Statistical Power Rates*

The SPOWER rates increase with the number of false null hypotheses increase in each procedure. In MHT procedures, the B-H procedure has the greatest SPOWER rates
and the Bonferroni procedure has the smallest. The Hochberg procedure and the Holm procedure rank the second and the third.

The TPOWER rates in the unadjusted procedure and the Bonferroni procedure are very stable, which are .80 and .54 respectively. With more false null hypotheses, the TPOWER rates increase in the Holm procedure, the Hochberg procedure, and the B-H procedure. Specifically, the TPOWER of the B-H procedure is the highest, the Hochberg procedure and the Holm procedure rank the second and the third respectively. Meanwhile, the TPOWER difference between the Holm procedure and the Hochberg procedure looks very small. The TPOWER in the Bonferroni procedure is the smallest.

With the number of false null hypotheses increases, the TDR in each procedure increases. The TDR is 1.0 with seven false null hypotheses. The TDR in the Bonferroni procedure is the greatest. The Holm procedure and the Hochberg procedure are the second and the third. The B-H procedure has the smallest TDR in MHT procedures. The unadjusted procedure has lower TDR than MHT procedures. Meanwhile, the differences of TDR among procedures become very small with more false null hypotheses.
CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

The results were discussed mainly based on the research questions which were introduced in Chapter one. Then, summary and conclusions were made following the discussion of the results in this study and recommendations were provided in terms of future research and practical implications in related areas.

Discussion of the Results

This Monte Carlo study mainly investigated the Type I error rates and statistical power rates of four MHT procedures in the balanced two-way and three-way factorial ANOVA. Meanwhile, the unadjusted alpha per test procedure was compared to MHT procedures. Specifically, four patterns of true null and false null effects in the two-way design and eight patterns for the three-way design were used to generate data with medium effect size of .50 and fixed sample size per cell (32 sample size per cell in the two-way design and 16 for the three-way design). Samplewise FWER, Testwise FWER, and FDR were used to evaluate the Type I error rates. Samplewise Power, Testwise Power, and TDR were adopted to indicate the statistical power rates. The discussion of the results was mainly based on the research questions in this study.

The first research question: How well do different MHT procedures control Type I error rates in terms of Samplewise FWER (SFWER), Testwise FWER (TFWER), and FDR in the balanced two-way and three-way factorial ANOVA?

The SFWER in the unadjusted alpha per test procedure is controlled at .05 with a single hypothesis test in both two-way and three-way factorial ANOVA. However, the SFWER inflates greatly with multiple hypothesis tests. For example, the SFWER is .0974
with two true null hypotheses and .1393 with three hypotheses in the two-way design. Also, it becomes .2913 with seven true null hypotheses, .2627 with six true null hypotheses, and .2217 with five true null hypotheses in the three-way design. Meanwhile, these SFWER are far beyond the liberal criterion of [.025, .075] (Bradley, 1978). Many researchers have emphasized the inflation of the Type I error rate when using the unadjusted alpha per test procedure in the context of MHT (Fletcher, Daw, & Young, 1989; Kromrey & Dickinson, 1995; Stevens, 1999; Schochet, 2008; Tabachinick & Fidell, 2001; Toothaker, 1993). That is, MHT procedures have been developed to help control the inflation of Type I error rates at the experimentwise or samplewise level. This study confirms their concerns about the inflation of Type I error rate in the unadjusted procedure in MHT. Therefore, researchers should avoid using the unadjusted alpha per test procedure in the context of MHT and may consider some MHT procedures.

Some researchers suggested using the experimentwise error rate (EWER) in factorial ANOVA (Maxwell & Delaney, 2000). In this current Monte Carlo simulation, one sample dataset is drawn to represent one experiment in each replication. With 20,000 trials, there are 20,000 experiments or sample datasets. In this sense, we can call it either EWER or SFWER, but we have chosen to use the more general term SFWER in this Monte Carlo study because it more accurately reflects the per-sample nature of the analysis desired. For example, multiple one-sample $t$-tests and multiple pairwise comparison can be performed in a sample but not constitute an experiment (Williams, Jones, & Tukey, 1999).
The Bradley (1978) criteria were used to measure the extent of SFWER departing from the nominal alpha level in each procedure in this study. Since researchers should make a balance between the Type I error and Type II error (Hochberg & Benjamini, 1990), the actual $p$ values should not be far from the nominal level, usually .05. That is, if $p$ values are much greater than .05, Type I error rates inflate greatly, then it is not necessary to evaluate power; if $p$ values are much less than .05, Type II error goes up and power decreases accordingly. Moreover, obtaining inflated Type I error seems more serious than having $p$ values much less than .05.

The Bonferroni procedure, the Holm procedure, and the Hochberg procedure are able to control the SFWER at accurate levels under the patterns of true null and false null effects in both two-way and three-way designs in terms of Bradley’s most stringent criterion (Bradley, 1978). However, the SFWER in the B-H procedure is not always within the liberal criterion of Bradley (1978) in this study. In fact, the SFWER in the B-H procedure were not as stable as the other three MHT procedures, that is, the SFWER of the B-H procedure inflate under certain conditions in terms of the liberal criterion in Bradley (1978). The B-H procedure control the FWER in a weak sense (Benjamini & Hochberg, 1995), the current study confirms this point.

There is an interesting finding about the SFWER in the B-H procedure. The distribution of the SFWER looks symmetrical. For example, in the two-way analysis, the SFWER is the greatest with two true null hypotheses; for one true null and three true null hypotheses, the SFWER are almost the same and lower than the SFWER with two true null (see Figure 1). In the three-way design, the SFWER is the greatest with four true null
hypotheses. The SFWER are almost the same for one true and seven true null hypotheses, two true and six true null hypotheses; and three true and five true null hypotheses. Meanwhile, the SFWER increase from seven true to four true, and decrease from four true to one true null (see Figure 7).

There are few studies that focused on the Type I error rates of the B-H procedure (Williams, Jones, & Tukey, 1999; Zhang, 2005), however, no research has been found that provided such interesting finding about the SFWER in the B-H procedure. The relationship between the Hochberg procedure and the B-H procedure was analyzed in Benjamini and Hochberg (1995), which proved that the ratio of $p$ values between the B-H and the Hochberg procedure is the largest at $i=(m+1)/2$ where $i$ is the $i$th test and $m$ is the number of hypothesis tests. For example, if there are seven tests and the ratio is the largest at the fourth test. Therefore, these analyses may be helpful to explain why the SFWER is the highest with four true null hypotheses in the three-way and with two true null in the two-way design in current study.

In the study of Kromrey and Dickinson (1995), the Type I error rates in the two-way and three-way balanced factorial design were investigated. With one true, two true, and three true null hypotheses in the two-way analysis, the unadjusted procedure has the Type I error rate of .05, .10, and .14, respectively; the Type I error rate in the Bonferroni procedure is .02, .03, and .05 respectively; in the Holm procedure, the Type I error rate is .03, .04, and .05; the Type I error rate is .03, .04, and .05, respectively in the Hochberg procedure. With four true, five true, six true, and seven true null hypotheses in the three-way design, the Type I error rate in the unadjusted procedure is .18, .22, .26, and .29,
respectively; .03, .04, .04, and .05 are in the Bonferroni procedure; .04, .04, .05, and .05 are in the Holm procedure and the Hochberg procedure. The SFWER in this study are consistent with the findings in Kromrey and Dickinson (1995) and the current results are more precise because of using larger iteration numbers (i.e. 20,000). Meanwhile, the Type I error rates for other three patterns, such as one true, two true, and three true null hypotheses, were not provided in Kromrey and Dickinson (1995). The current study provides more complete results than theirs.

In the study of Morikawa, Terao, and Iwasaki (1996), the Type I error rates of several MCPs were investigated for multiple groups, for example, when there are three groups. The Type I error rates result from the Bonferroni procedure, the Holm procedure, and the Hochberg procedure are .046, .046, and .047 under the overall null hypotheses, which are close to the results in the current study: .0476, .0476, and .0478 with three true null hypotheses in the two-way balanced design.

In the study of Williams, Jones, and Tukey (1999), FWER of the Bonferroni procedure, the Hochberg procedure, and the B-H procedure were investigated based on 48 uncorrelated $t$-tests and 1128 pairwise comparisons in a simulation. As a result, the Bonferroni procedure and the Hochberg procedure can maintain the Type I error rate at an acceptable level in all conditions of the two scenarios in their study. “The B-H procedure fails to maintain the familywise error rate (which it was never designed to maintain)” (p. 58) under certain conditions, for example, under the moderate and large effect sizes. In this current study, the inflation of the SFWER of the B-H procedure under some patterns was also found, even with equal effect sizes across all conditions.
Meanwhile, the study of Smith et al. (2002) found that the inflation of Type I error rate occur in the unadjusted alpha procedure and the Bonferroni procedure can reduce the Type I error rate inflation in factorial ANOVA.

The TFWER is the average FWER based on true null hypotheses of all the tests in the simulation. The TFWER in the unadjusted procedure is controlled at .05 under each pattern in this study. The Bonferroni procedure also has the constant TFWER which is about .0166 and .007 in the two-way and three-way design respectively. The TFWER rates increase with the number of the true null hypotheses decreases in the Holm procedure, the Hochberg procedure, and the B-H procedure. In the two-way design, the average TFWER in the Holm procedure is about .0254, the average TFWER in the Hochberg procedure is .0274, and the average TFWER in the B-H procedure is .03. In the three-way design, the average TFWER rates are .0091, .0093, and .0191, respectively. No research has been found to present the TFWER of several MHT procedures; this study provides a new perspective to evaluate MHT procedures. For example, the TFWER result from the Holm procedure, the Hochberg procedure, and the B-H procedure indicate that they are “sequentially rejective” procedures.

The FDR is 1.0 under the pattern of all true null hypotheses in both two-way and three-way designs. Therefore, the FDR can’t be controlled. However, with more false null hypotheses, FDR becomes different in each procedure. As Benjamini and Hochberg (1995) stated, “for independent test statistics and for any configuration of false null hypotheses, the procedure controls the FDR at \( q^* \)” (p. 293). Therefore, the FDR is controlled at \( q^* \) because \( E(Q) \leq \frac{m_0}{m} q^* \leq q^* \). Meanwhile, the \( q^* \) is set to equal to \( \alpha (.05) \)
when comparing the B-H procedure and the Hochberg procedure in the study of Benjamini and Hochberg (1995). Therefore, the FDR is controlled within .05. In this current study, for example, most FDR in the Bonferroni procedure, the Holm procedure, and the Hochberg procedure are below .05 except for six true null hypotheses in the three-way design, which indicates that the FWER procedures are able to control the FDR in a more conservative way. With five and six true null, the FDR in the B-H procedure are greater than .05. The FDR are controlled within .05 in other patterns. Meanwhile, with more false null hypotheses, the FDR in each procedure becomes smaller, which is consistent with the above formula of $E(Q) \leq \frac{m_0}{m} q^* \leq q^*$. 

Williams, Jones, and Tukey (1999) also investigated the FDR of the Bonferroni procedure, the Hochberg procedure, and the B-H procedure in the context of 48 uncorrelated $t$-tests and 1128 pairwise comparison. The FDR of three procedures were obtained for different effect size: .001, .30, 1.0, 3.0, and 5.0. Since different conditions and designs were used in their study, their results wouldn’t be exactly same as the current study. However, some results may be useful. For example, the FDR are less than .05 in all procedures. More specifically, the FDR in the B-H procedure is greater than the Bonferroni procedure and the Hochberg procedure with varied effect size. The FDR results from the unadjusted procedure are greater than MHT procedures. In the study of Schochet (2008), the FDR of the B-H procedure were calculated based on multiple $t$-tests. The results were very similar as the current study which confirms the generalizability of current study because factorial ANOVA was used in this study.
The second research question: How well do different MHT procedures show statistical power rate in terms of Samplewise Power, Testwise Power, and TDR in the balanced two-way and three-way factorial ANOVA?

It should be noted that the Type I error rate results from the unadjusted alpha per test procedures are actually inflated greatly with more than one hypothesis test, it is not necessary to compare the power rates of this procedure with MHT procedures in the context of MHT (Keselman, Cribbie, & Holland, 2002). Specifically, in terms of SPOWER and TPOWER, the Bonferroni procedure is the smallest and the B-H procedure is the greatest. The Hochberg procedure and the Holm procedure rank the second and the third respectively.

In the study of Kromrey and Dickinson (1995), the statistical power rates in the two-way and three-way balanced factorial design were investigated. The any effect power and average per effect power were provided, which correspond to the SPOWER and TPOWER in current study. Because of using different effect sizes for effects and different sample size per cell, the results are not exactly same as the current study. For example, with three false null hypotheses and 20 sample size per cell in the two-way analysis in Kromrey and Dickinson (1995), the any effect power rates are .66, .66, and .67 in the Bonferroni procedure, the Holm procedure, and the Hochberg procedure respectively; the average per effect power are .45, .50, and .51. While in the current study, the SPOWER rates are .9578, .9578, and .9629 and the TPOWER rates are .6535, .7474, and .7622. But the Hochberg procedure is more powerful than the Bonferroni procedure and the Holm procedure in both studies.
In the study of Morikawa, Terao, and Iwasaki (1996), the statistical power rates of several MCPs were investigated in multiple groups, for example, three groups and sample size of 50. The any-pair power and all-pairs power (they called it all-pairs power but they described it as the average power) correspond to the SPOWER and TPOWER in current study. With effect size of .50, the any-pair power rates are all 1.0 in the Bonferroni procedure, the Holm procedure, and the Hochberg procedure. In the current study, the SPOWER rates of the three procedures are .9578, .9578, and .9629. The all-pairs power rates are .496, .725, and .728 while the TPOWER rates are .6535, .7474, and .7622 in the current study. Therefore, these two studies are consistent.

Williams, Jones, and Tukey (1999) used the average power rate in their study. As they stated, “for both uncorrelated and pairwise families, the B-H technique results in greater power than that for the Hochberg or Bonferroni procedures” (p. 58). In the study of Schochet (2008), multiple t-tests were used to investigate the statistical power rates of the Bonferroni procedure, the Holm procedure, and the B-H procedure. As a result, the B-H procedure was proved to have more power advantage than other procedures. In the study of Klockars and Hancock (1992), the statistical power rates for multiple groups (four groups and five groups) were investigated using several MHT procedures. The Hochberg procedure shows more power than the Holm procedure in their analysis.

Some researchers pointed out that the Bonferroni procedure is conservative in terms of power (Hochberg, 1988; Hochberg & Benjamini, 1990; Holm, 1979; Shaffer, 1995; Schochet, 2008). The Holm procedure is more powerful than the Bonferroni procedure (Holm, 1979, Hochberg, 1988; Hochberg & Benjamini, 1990; Kromrey &
Dickinson, 1995; Schochet, 2008). The Hochberg procedure is more powerful than the Bonferroni and the Holm procedure (Hochberg, 1988; Kromrey & Dickinson, 1995). Meanwhile, it should be noted that the power differences in the Holm procedure and the Hochberg procedure are actually small (Olejnik, et al. 1997). The B-H procedure has power advantage over other MHT procedures especially when there are many known false null hypotheses (Benjamini & Hochberg, 1995; Keselman, Cribbie, & Holland, 2002; Schochet, 2008). The results in this Monte Carlo study confirm previous findings.

Besides SPOWER, the TPOWER is another indicator to evaluate the power. In this study, the SPOWER of MHT procedures seem clustered together (see Figure 4 and Figure 10), the SPOWER difference in each MHT procedure is very small. TPOWER provides a clear idea of each MHT procedure (see Figure 5 and Figure 11). Even though TDR is just 1-FDR, the TDR is a new power indicator in this study. When all hypotheses are false null, the TDR is 1.0 in both two-way and three-way designs. With true null hypotheses added in, the TDR are less than 1.0 in all procedures. In fact, with more false null hypotheses, the differences in TDR in MHT procedures become very small. No studies that have been found to investigate the TDR of MHT procedures. Verhoevne, Simonsen, and McIntyre (2005) stated that “a natural companion to the FDR is the false nondiscovery rate (FNR), or the expected proportion of nonrejections that are incorrect” (p. 646). In this study, with more false null hypotheses, the TDR differences in MHT procedures become very small. For example, when there are six false null hypotheses, the TDR is almost the same in each procedure.
Conclusions

This Monte Carlo study investigated the Type I error rates and statistical power rates of four MHT procedures in the balanced two-way and three-way factorial ANOVA. There are no studies that have been found to investigate the FWER in the B-H procedure and the FDR in the Bonferroni procedures and modified Bonferroni procedures in factorial ANOVA. Therefore, several useful conclusions are made from this Monte Carlo study.

First, the Type I error rate (SFWER in this study) in the unadjusted procedure inflates greatly with multiple hypothesis tests. However, some MHT procedures have advantages in this study which are better able to control the inflation of Type I error with multiple tests. Therefore, MHT procedures are useful for researchers in the context of MHT. Meanwhile, the unadjusted procedure should be avoided in MHT.

Second, the Bonferroni procedure is able to control the SFWER at an accurate level with conservative power in the balanced factorial ANOVA under several patterns. Specifically, in the three-way design, one SFWER in the Bonferroni procedure is in the stringent criterion, one SFWER in the intermediate criterion, and two in the liberal criterion. The other three SFWER are out of the liberal criterion. In the two-way design, one SFWER is in the stringent criterion, one SFWER is in the intermediate criterion, and the other one is beyond the liberal criterion (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999). The TFWER in the Bonferroni procedure is controlled at a constant level. There is one FDR in the Bonferroni procedure is greater than .05 in the three-way design, the others are less than .05. The FDR in the two-way analysis are both less than .05.
Third, the Holm procedure is proved to be able to control the SFWER in terms of Bradley (1978) criteria in almost all patterns with better power than the Bonferroni procedure in this study. In the three-way design, only one SFWER in the Holm procedure is beyond the liberal criterion, two SFWER are in the liberal criterion, two SFWER are in the intermediate criterion, and two are in the stringent criterion. In the two-way design, one SFWER is in the liberal criterion, one is in the intermediate criterion, and one is in the stringent criterion (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999). The TFWER in the Holm procedure indicates that it is a sequentially rejective procedure. With less true null hypotheses, the TFWER become greater. Only one FDR in the Holm procedure is greater than .05 and others are less than .05 in the three-way design. In the two-way analysis, one FDR is greater than .05 and one is less than .05.

Fourth, the Hochberg procedure can successfully control the SFWER at an accurate level with multiple hypotheses tests in the balanced factorial ANOVA. In the three-way design, all SFWER in the Hochberg procedure are within the liberal criterion. Specifically, three SFWER are in the liberal criterion, two SFWER are in the intermediate criterion, and two SFWER are in the stringent criterion. In the two-way design, one SFWER is in the intermediate criterion and two SFWER are in the intermediate criterion (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999). The TFWER in the Hochberg procedure indicates that it is a sequentially rejective procedure. With less true null hypotheses, the TFWER become greater. Only one FDR in the Hochberg procedure is greater than .05 and others are less than .05 in the three-way design. In the two-way analysis, one FDR is greater than .05 and one is less than .05.
Fifth, the SFWER in the B-H procedure inflate under some patterns in this study. Specifically, in the three-way design, three SFWER are beyond the liberal criterion, two are in the intermediate criterion, and two are in the liberal criterion. In the two-way design, two SFWER are in the intermediate criterion and one is in the stringent criterion (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999). The TFWER in the B-H procedure indicates that it is a sequentially rejective procedure. With less true null hypotheses, the TFWER become greater. Four FDR in the B-H procedure are less than .05 in the three-way design and two are greater than .05. In the two-way design, one FDR is greater than .05 and one is less than .05.

Sixth, the B-H procedure is proved to be more powerful than other MHT procedures. The Bonferroni procedure is less powerful than other MHT procedures. The Hochberg procedure and the Holm procedure rank the second and the third in terms of SPOWER and TPOWER in this study.

In summary, the Type I error inflation occur in the unadjusted procedure in MHT. Some MHT procedures can be used to better control the Type I error inflation in MHT. The Hochberg procedure has relative advantages than other MHT procedures in terms of controlling SFWER at an accurate level (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999). Meanwhile, it is slightly more powerful than the Bonferroni and the Holm procedure. The power difference between the Hochberg procedure and the B-H procedure is very small in this study. Although the B-H procedure is the most powerful in this study, the SFWER in this procedure inflate in some conditions and some FDR are also
greater than .05 (even it aims to control the FDR below .05). That is, the power advantage of the B-H procedure comes at the cost of a little bit inflated SFWER.

Based on the current study, not only does the B-H procedure control the FDR, but also the FWER procedures are able to control FDR in a more conservative manner (FDR in FWER procedures were less than the B-H procedure). Meanwhile, the B-H procedure can control FWER in a weak sense (when there are all true null). TFWER, TPOWER, and TDR provide new aspects to evaluate MHT procedures. For example, TFWER in MHT procedures indicates the characteristic of “sequentially rejective” in the Holm, the Hochberg, and the B-H procedure. The TPOWER provides a clear idea about MHT procedures in terms of the power in each test because the SPOWER differences of MHT procedures are small. The FDR and TDR are complement to each other, which enable TDR become a new indicator to assess MHT procedures, although the information that provided by TDR was not much in this study. Finally, as long as the assumption held, the Hochberg procedure worked best in current study overall. If the assumption is not met, the Holm procedure is recommended.

Recommendations

In previous literature, MHT procedures were discussed in the context of mainly in multiple comparisons of one-way ANOVA, in addition, only the Bonferroni procedure and some modified Bonferroni procedures were included. But the B-H procedure was included in this study. Meanwhile, the Type I error rate and statistical power rate were evaluated in several perspectives which can provide researchers a comprehensive picture.
about MHT procedures. Specifically, this study can provide some recommendations in terms of (a) future research, and (b) practical implications.

**Future Research**

This Monte Carlo study analyzed the Type I error rate and statistical power rate of MHT procedures in the balanced factorial ANOVA, there are certain findings, as well as some limitations, that suggest future research.

The unique characteristic of the balanced factorial ANOVA is that all hypothesis tests are independent, which allow some MHT procedures that have the assumption of independence (e.g. the Hochberg and the B-H procedures) to be included in current study. That is one reason why the balanced factorial ANOVA was used as the context to evaluate MHT procedures. Admittedly, the unbalanced factorial designs are also very common; therefore, future research can focus on the unbalanced designs in which the effects are dependent or correlated.

There are several types of statistical power: any-pair power, all-pair power, and average per-pair power (Kirk, 1995). Only the SPOWER and TPOWER were used in this study. In future research, the power can be evaluated from another point of view. For example, the Type II error can be calculated in a Monte Carlo simulation, power can be obtained through one minus Type II error. Although TDR is not exactly the same as statistical power, it may become an indicator when FDR is included because they are complement to each other.

Researchers should have a better understanding about FDR. The FDR is set to equal alpha level in the study of Benjamini and Hochberg (1995). There are not many
studies that have been found to investigate a good value for FDR (Benjamini & Hochberg, 1995; Williams, Jones, & Tukey, 1999). The Type I error should be controlled at .05, but for FDR, it is not clear that the smaller is the better, or it also has to be controlled at .05. Meanwhile, the B-H procedure shows power advantage in current study, the relationship between the FDR and the power can become a research question for future research. Also, the symmetrically distributed SFWER in the B-H procedure can become a future topic. For example, why and when the highest SFWER in the B-H procedure occur?

The Holm procedure, the Hochberg procedure, and the B-H procedure are all sequentially rejective procedures. More specifically, the B-H procedure is regarded as linear rejective procedure and the Holm and the Hochberg procedures are nonlinear. However, the linear procedure lead to nonlinear SFWER and the nonlinear procedures have relatively linear SFWER (see Figure 7). Future research can focus on these interesting issues.

All hypothesis tests were regarded as one family in this study, future study may treat them as separate families since there are different opinion about the family in factorial ANOVA. Also, the sample size per cell was fixed in this study; future study may compare the Type I error rate and statistical power rate with different sample size per cell. The assumptions were held in this study, future study may consider the situations when assumptions are violated. Although some researchers investigated the consequences under the assumption violations in factorial ANOVA (Hsiung & Olejnık, 1996; Olejnık
& Algina, 1985), no studies have been found to include MHT procedures in factorial ANOVA when assumptions are violated.

The B-H procedure is a relatively new procedure and several kinds of adaptive B-H procedures are available for researchers. Monte Carlo simulation was used to achieve the research purposes in this study. Recently, the properties of FDR have been investigated in a mathematical way, for example, the proof of the application of FDR in the dependent hypothesis tests was illustrated mathematically in the study of Benjamini and Yekutieli (2001). The Bayesian probability was used to find the $q^*$ value for FDR (Sarkar, 2007; Storey, 2002, 2003, 2007). Therefore, if the properties of FDR are interested, additional mathematical analysis will be another option for future studies.

Practical Implications

The Type I error rate inflation in MHT not only exists in the one-way ANOVA, but also in the factorial ANOVA. To control the Type I error inflation, the current study provides researchers with some practical implications in terms of recommending different MHT procedures under certain conditions. The selection of MHT procedures is mainly based on the Type I error rate and statistical power rate (Kirk, 1995; Toothaker, 1993). In addition, the ease of application and interpretation is another factor from a practical point of view (Keren & Lewis, 1993; Olejnik, Li, Supattathum, & Huberty, 1997). As Brown and Russell (1997) stated, “multiple hypothesis tests are sometimes performed to avoid very complex analyses whose results would be difficult to communicate” (p. 2511).

The Unadjusted Alpha Per Test Procedure
It is obvious to see the Type I error rate inflates greatly when using the unadjusted alpha per test procedure in this study. Therefore, it is necessary for researchers to think about controlling the Type I error rate using MHT procedures in practice. In fact, many researchers have made arguments in terms of controlling the Type I error inflation in the context of MHT (e.g., Games, 1971; Ryan, 1959; Stevens, 1999). The unadjusted procedure should be avoided in the context of MHT since MHT procedures have advantages in controlling the Type I error inflation.

The Bonferroni Procedure

Although the Bonferroni procedure has the smallest power rate in this study, the performance of this procedure is stable and consistent under all patterns of true null and false null effects. Also, it is easily calculated (even by hand) and understood by readers who are not statistics majors. Therefore, the Bonferroni procedure is a good choice for researchers who are dealing with the dataset in which has fewer hypothesis tests. Generally, the Bonferroni procedure is regarded as the simplest and the best known MHT procedure (Olejnik, et al. 1997). Meanwhile, there is no restricted assumption under this procedure.
The Holm Procedure

The Holm procedure requires no assumptions (Olejnik, et al. 1997; Schochet, 2008). It has been proved to be more powerful than the Bonferroni procedure in current study; meanwhile, the SFWER control in this procedure is more accurate than the Bonferroni procedure. Therefore, it is a better choice than the Bonferroni procedure.

The Hochberg procedure

The Hochberg procedure is slightly more powerful than the Bonferroni procedure and the Holm procedure (Hochberg, 1988; Kromrey & Dickinson, 1995); however, it only can be applied in the independent hypotheses tests (Olejnik, et al. 1997; Schochet, 2008). Therefore, as long as the assumption is met, the Hochberg procedure can bring slightly more power than the Holm procedure and the Bonferroni procedures (Kromrey & Dickinson, 1995). As Holland and Copenhaver (1987) suggested, a modified Bonferroni procedure should be used when otherwise the original Bonferroni procedure is the choice. Meanwhile, the SFWER control in the Hochberg procedure is more accurate than other MHT procedures in terms of criteria (Bradley, 1978; Brooks, Barcikowski, & Robey, 1999) in current study.

The B-H Procedure

Although FDR is a nice way to think about Type I error rate, some issues in FDR are not clear to researchers, for example, the relationship between FDR and power. Meanwhile, the power advantage of the B-H procedure comes at the cost of a little bit inflated SFWER, the use of the B-H procedure should be carefully considered.

Final Recommendations
In fact, the current study proved that the FWER procedures are able to control FDR in a conservative manner. That is, the B-H procedure is not the only method in terms of controlling FDR. The B-H procedure can control the FWER only in a weak sense. Therefore, FWER procedures can be recommended instead of using the B-H procedure. Although some researchers concerned about the FWER, as Storey (2003) stated, “The FWER offers an extremely strict criterion which is not always appropriate” (p. 2014). Stevens (1996) already suggested using a liberal level of alpha, for example, .10 or .15 in MHT, which is sort of equivalent to FDR with its “weak control” (Benjamini & Hochberg, 1995; p. 291). That is, the liberal alpha level can lead to powerful results instead of using FDR approaches. Overall, as long as the assumption is held, the Hochberg procedure worked best in current study. If the assumption is not met, the Holm procedure is recommended.

It is necessary to noted here that several indicators that were used in current study also have some practical implications. For example, the testwise criteria (TFWER and TPOWER) can provide a clearer idea about MHT procedures than only using samplewise criteria (SFWER and SPOWER). For example, the TFWER indicates that the Holm procedure, the Hochberg procedure, and the B-H procedure are sequentially rejective procedures because the TFWER in these procedures increases with less true null hypotheses. The TPOWER indicates that which MHT procedure has the greatest power in each hypothesis test.

The generalizability of this study has been guaranteed by using the balanced factorial ANOVA and employing representative patterns of true null and false null
effects. For example, the balanced factorial designs have independent hypothesis tests which allow the Hochberg and the B-H procedure to be used in this study. The number of patterns of true null and false null effects was emphasized instead of focusing on which effect is true null and false null, that is, if there are fixed number of true null and false null effects no matter which effect is true and false null, the performance of MHT procedures will be consistent as current study. Moreover, the current results are consistent with the study of Schochet (2008) which has a different kind of research design. In this sense, researchers can directly use the results in current study for their analyses if independent hypothesis tests are included.

Finally, although researchers have different opinions about adjusting alpha or not, they agreed that MHT procedures are useful statistical tools (Hewes, 2003; O’Keefe, 2003; Tukey, 1991). The usage of MHT procedures depends on whether researchers adjusting alpha or not. It is a philosophical issue and researchers have to make their real decisions. If the alpha should be adjusted, then how to adjust the alpha relates to the definition of family and the focal unit of Type I error (which kind of Type I error should be controlled). As Tukey (1991) suggested, there is no unique answer toward controlling the Type I error inflation.
REFERENCES


Hsiung, T., & Olejnik, S. (1994, April). *Type I error rates and statistical power for the James Second-Order test and the univariate F test in two-way fixed-effects ANOVA*
models under heteroscedasticity and/or nonnormality. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.


APPENDIX A: MULTIPLE HYPOTHESIS TESTING PROCEDURES

Several alpha-adjustment MHT procedures are used in testing multiple hypotheses. The following section will illustrate four alpha-adjustment MHT procedures: The Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the B-H procedure. In addition, the unadjusted alpha per test procedure is included to compare with the alpha-adjustment MHT procedures.

A three-way ANOVA example is adapted from Neter, Kutner, Nachtsheim, and Wasserman (1996). The factors are gender of subject (factor $A$), body fat of subject (factor $B$), and smoking history (factor $C$). The dependent variable is the exercise tolerance ($Y$). Each factor has two levels. Finally, a total of seven hypotheses are tested simultaneously: three main effects ($A$, $B$, and $C$), three two-way interactions ($AB$, $AC$, and $BC$), and one three-way interaction ($ABC$).
Table 8

*Three-way ANOVA Output*

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>df</th>
<th>Mean-Square</th>
<th>F-Ratio</th>
<th>p values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>168.584</td>
<td>1</td>
<td>168.584</td>
<td>18.059</td>
<td>0.001</td>
</tr>
<tr>
<td>B</td>
<td>242.570</td>
<td>1</td>
<td>242.570</td>
<td>25.984</td>
<td>0.000</td>
</tr>
<tr>
<td>C</td>
<td>74.384</td>
<td>1</td>
<td>74.384</td>
<td>7.968</td>
<td>0.012</td>
</tr>
<tr>
<td>AB</td>
<td>13.650</td>
<td>1</td>
<td>13.650</td>
<td>1.462</td>
<td>0.244</td>
</tr>
<tr>
<td>AC</td>
<td>11.070</td>
<td>1</td>
<td>11.070</td>
<td>1.186</td>
<td>0.292</td>
</tr>
<tr>
<td>BC</td>
<td>76.454</td>
<td>1</td>
<td>76.454</td>
<td>8.190</td>
<td>0.011</td>
</tr>
<tr>
<td>ABC</td>
<td>1.870</td>
<td>1</td>
<td>1.870</td>
<td>0.200</td>
<td>0.660</td>
</tr>
<tr>
<td>Error</td>
<td>149.367</td>
<td>6</td>
<td>9.335</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Unadjusted Alpha Per Test Procedure*

Each *p* value is compared to the nominal alpha (α) level (.05) in this method. All three main effects (*A*, *B*, and *C*) and one two-way interaction (*BC*) are significant at .05. For example, the Smoking History (Factor *C*) had a *p* = .012 which is less than the .05 nominal level of significance, and therefore the null hypothesis of equal means for that main effect is rejected.

*The Bonferroni Procedure*

The Bonferroni Inequality (cited in Stevens, 1996) states that if *n* hypotheses are separately tested with Type I error rates of $\alpha_1, \alpha_2, ..., \alpha_n$, then overall

$\alpha \leq \alpha_1 + \alpha_2 + ... + \alpha_n$. Consequently, if each hypothesis is tested at the same alpha level
$(\alpha_i)$, the level of $\alpha/n$ should be used for each individual test. The null hypothesis is rejected if the individual $p$ value is less than $\alpha/n$, therefore the overall $\alpha$ is controlled, such that overall $\alpha \leq n\alpha_i$. For the Neter et al. (1996) example above, each $p$ value is evaluated at $\alpha/7$ and if an obtained $p$ value is less than .05/7, or .007, the null hypothesis is rejected. As a result, only the main effects $A$ and $B$ are significant, because their $p$ values are less than the Bonferroni alpha of .007.

In addition, we can conduct the Bonferroni procedure in the logic of adjusting $p$ values. That is, each obtained $p$ value can be multiplied by 7, and then compared to .05. This is the rationale behind the R program (see http://www.r-project.org/) which adjusts $p$ values instead of the alpha level. It requires that each adjusted $p$ value is compared to .05. If the adjusted $p$ value is less than .05, the null hypothesis is rejected; otherwise, the null is not rejected. Meanwhile, since the adjusted $p$ values represent the probability and cannot be larger than 1.0. Therefore, all adjusted $p$ values over 1.0 are set to 1.0.
Table 9

Using the Bonferroni Procedure

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$p$-values</th>
<th>Adjusted Alpha</th>
<th>Adjusted $p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>0.001</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>$B$</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>$C$</td>
<td>0.012</td>
<td>0.007</td>
<td>0.084</td>
</tr>
<tr>
<td>$AB$</td>
<td>0.244</td>
<td>0.007</td>
<td>1.000</td>
</tr>
<tr>
<td>$AC$</td>
<td>0.292</td>
<td>0.007</td>
<td>1.000</td>
</tr>
<tr>
<td>$BC$</td>
<td>0.011</td>
<td>0.007</td>
<td>0.077</td>
</tr>
<tr>
<td>$ABC$</td>
<td>0.660</td>
<td>0.007</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note: ** represents adjusted $p$ values $\geq 1.0$.

The Holm Procedure

A more powerful and logical sequentially rejective Bonferroni procedure was proposed in Holm (1979). Several researchers illustrated this step-down method (Kromrey & Dickinson, 1995; Shaffer, 1995; Schochet, 2008), the $p$ values for the $n$ hypotheses being tested are ordered from the smallest to largest ($p_1 \leq p_2 \ldots \leq p_n$). The smallest $p$ value ($p_1$) is tested at $\alpha / n$, if $p_1 > \alpha / n$; all the $n$ hypotheses are not significant, otherwise, the null hypothesis is rejected and the next hypothesis is considered. The second smallest $p$ value ($p_2$) is tested at $\alpha / (n-1)$; if $p_2 > \alpha / (n-1)$, all subsequent tests are nonsignificant, otherwise the null is rejected and the next smallest $p$
value \( p_3 \) is considered. The test continues until \( p_i > \alpha / (n - i + 1) \), where \( i \) is the \( i \)th smallest \( p \) value.

In this example, the smallest \( p \) value of factor \( B \) is close to 0.0 and it is less than \( 0.05 / (7-1+1) \) or .007; therefore, the null hypothesis is rejected. The second smallest \( p \) value of factor \( A \) is close to .001 and it is less than \( 0.05 / (7-2+1) \) or .008, so that the null hypothesis is rejected. The third smallest \( p \) value is .011 from the interaction \( BC \), and it is greater than \( 0.05 / (7-3+1) \) or .01, therefore, the hypothesis is not rejected and all the subsequent hypotheses are nonsignificant. That is, only the main effects (\( A \) and \( B \)) are significant.

In addition, the logic of adjusting \( p \) values can be applied. The smallest \( p \) value \( (p_1) \) is multiplied by \( (7-1+1) \) or 7 and the adjusted \( p \) value is less than .05, and the null is rejected. The second smallest \( p \) value is multiplied by 6. The adjusted \( p \) value is less than .05, and the null is rejected. The third smallest \( p \) value (.011) is multiplied by 5. The adjusted \( p \) value .055 is greater than .05. Therefore, both main effects \( B \) and \( A \) are significant, others are nonsignificant.
Table 10

*Using The Holm Procedure*

<table>
<thead>
<tr>
<th>$i$th p-value</th>
<th>$n-i+1$ Hypothesis</th>
<th>$p$-values</th>
<th>Adjusted Alpha</th>
<th>Adjusted $p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.001</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.011</td>
<td>0.010</td>
<td>0.055</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.012</td>
<td>0.013</td>
<td>0.048</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.244</td>
<td>0.017</td>
<td>0.732</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.292</td>
<td>0.025</td>
<td>0.584</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.660</td>
<td>0.050</td>
<td>0.660</td>
</tr>
</tbody>
</table>

*The Hochberg Procedure*

Hochberg (1988) proposed a step-up modified Bonferroni procedure. Some researchers summarized this procedure (Kromrey & Dickinson, 1995; Shaffer, 1995; Schochet, 2008). Unlike Holm’s method, $p$ values of $n$ tests are ordered from the largest to the smallest in Hochberg’s method ($p_n \geq p_{n-1} \geq \ldots \geq p_2 \geq p_1$). The largest $p$ value ($p_n$) is tested at $\alpha / 1$, if $p_n < \alpha$, all the subsequent tests are rejected. Otherwise, the next largest $p$ value ($p_{n-1}$) is tested at $\alpha / 2$, if $p_{n-1} < \alpha / 2$, all the subsequent hypotheses are rejected. Otherwise, the third largest $p$ value is examined. Generally, if $p_i < \alpha / (n - i + 1)$, $i$ is the $i$th $p$ value, all the subsequent hypotheses are rejected. In fact, the critical values
in this procedure are the same as that in Holm’s method; the difference is the test sequence of \( p \) values (Kromrey & Dickinson, 1995).

For this example, the largest \( p \) value \( (p_7) \) is .660 and it is greater than \( .05/(7-7+1) \) or .05, therefore the null is not rejected. The next \( p \) value \( (p_6) \) is examined, obviously, 0.292 is greater than \( .05/(7-6+1) \), or .025, the null is not rejected. The next \( p \) value \( (p_5) \) .244 is greater than \( .05/(7-5+1) \), or .017, the null is not rejected. Actually, the null is not rejected until the \( p \) value of the main effect \( C(p_4) \) is tested, it is less than \( .05/(7-4+1) \) or .013, therefore, the subsequent hypotheses are rejected. The main effect \( A, B, C \) and interaction term \( BC \) are significant.

In addition, the logic of adjusting \( p \) values can be applied. The largest \( p \) value \( (p_7) \) is multiplied by \( (7-7+1) \) or 1, that is, the adjusted \( p \) value .660 is greater than .05, the null is not rejected. The next \( p \) value \( (p_6) \) is multiplied by \( (7-6+1) \) or 2, that is, .584 is greater than .05, the null is not rejected. The adjusted \( p \) value of the main effect \( C \) is less than .05; therefore, the subsequent hypotheses are rejected. The main effect \( A, B, C \) and interaction term \( BC \) are significant.
Table 11  

Using The Hochberg Procedure

<table>
<thead>
<tr>
<th>$i^{th}$ p-value</th>
<th>$n-i+1$</th>
<th>Hypothesis</th>
<th>p-values</th>
<th>Adjusted Alpha</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>ABC</td>
<td>0.660</td>
<td>0.050</td>
<td>0.660</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>AC</td>
<td>0.292</td>
<td>0.025</td>
<td>0.584</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>AB</td>
<td>0.244</td>
<td>0.017</td>
<td>0.732</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>C</td>
<td>0.012</td>
<td>0.013</td>
<td>0.048</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>BC</td>
<td>0.011</td>
<td>0.010</td>
<td>0.055</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>A</td>
<td>0.001</td>
<td>0.008</td>
<td>0.006</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>B</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The Benjamini-Hochberg Procedure

Benjamini and Hochberg (1995) developed the B-H procedure which controls the FDR within the $q^*$ level where $q^*$ equals $\alpha$. The steps of B-H method are: the $p$ values of the $n$ tests are ordered from the smallest to the largest ($p_1 \leq p_2 \leq \ldots p_n$), find the largest $i^{th}$ $p$ value which satisfies that $p_i \leq (i/n)^*\alpha$, reject all hypotheses from the first to the $i^{th}$ hypothesis (Benjamini & Hochberg, 1995; Cai, 2006; Keselman, Cribbie, & Holland, 2002; Schochet, 2008).

In this example, the first $p$ value ($p_1$) is less than $.05^*(1/7)$ or .007, the second $p$ value ($p_2$) is less than $.05^*(2/7)$ or .014, the third $p$ value ($p_3$) .011 is less than $.05^*(3/7)$ or .021, the fourth $p$ value ($p_4$) .012 is less than $.05^*(4/7)$ or .029, the fifth $p$ value ($p_5$)
.244 is greater than the adjusted alpha .036 (.05*5/7). Therefore, the first four hypotheses are rejected, that is, three main effects (A, B, and C) and one two-way interaction (BC) are significant.

In addition, the logic of adjusting p values can be applied, let the p value multiply \((n/i)\). For example, the first p value \((p_1)\) is multiplied by 7 (7/1), the adjusted p value is less than .05. The second p value \((p_2)\) is multiplied by 7/2. As a result, the adjusted p values for the second, the third, and the fourth hypothesis are all less than .05. Therefore, the first four hypotheses are significant.

<table>
<thead>
<tr>
<th>(i)th p-value</th>
<th>Hypothesis</th>
<th>(p)-values</th>
<th>Adjusted Alpha</th>
<th>Adjusted (p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>0.001</td>
<td>0.014</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td>BC</td>
<td>0.011</td>
<td>0.021</td>
<td>0.026</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>0.012</td>
<td>0.029</td>
<td>0.021</td>
</tr>
<tr>
<td>5</td>
<td>AB</td>
<td>0.244</td>
<td>0.036</td>
<td>0.342</td>
</tr>
<tr>
<td>6</td>
<td>AC</td>
<td>0.292</td>
<td>0.043</td>
<td>0.341</td>
</tr>
<tr>
<td>7</td>
<td>ABC</td>
<td>0.660</td>
<td>0.050</td>
<td>0.660</td>
</tr>
</tbody>
</table>

**Table 12**

*Using The B-H Procedure*
A Small Summary

The results and conclusions from MHT approaches are different. The main effects \((A\ and\ B)\) are significant in the Bonferroni procedure and the Holm procedure. The Hochberg procedure and the B-H procedure produce four significant effects. It is necessary to mention that the sequence of \(p\) values does affect the conclusion in the Holm procedure and the Hochberg procedure, although the critical values that are used in the two procedures are the same (Kromrey & Dickinson, 1995). The Holm procedure begins with the smallest \(p\) value and the test stops when the adjusted \(p\) value is greater than .05. In the example above, it stops at the third \(p\) value, .055 which is greater than .05. Even though the fourth \(p\) value (.048) is less than .05, the conclusion is that only the first two \(p\) values are significant. In the Hochberg procedure, the largest \(p\) value is first tested, the test stops when a \(p\) value is found to be smaller than .05. In this example, the fourth \(p\) value .048 is less than .05; therefore, all the subsequent \(p\) values are significant even though the next \(p\) value .055 is greater .05. That is, the three main effects \((A, B,\) and \(C)\) and interaction term \((BC)\) are significant.

In addition, different assumptions are underlying each MHT procedure. For example, both the Bonferroni procedure and the Holm procedure are recommended as suitable general-purpose methods in MHT because of no constraining assumptions (Holland & Copenhaver, 1987; Schochet, 2008). The Hochberg procedure shows more relative power than other procedures (e.g., the Bonferroni procedure and the Holm procedure). However, this procedure only works when the test statistics are independent (Holland & Copenhaver, 1987). The B-H procedure is proved to be suitable for
independent tests (Benjamini & Hochberg, 1995), and then it is also proved to be
workable under positively dependent tests (Benjamini & Yekutieli, 2001). Generally, the
B-H procedure requires that the number of total rejections should be greater than zero.
### APPENDIX B: R CODE FOR TWO-WAY BALANCED DESIGN

```R
### First, remove memory in the computer and set the seed for the generation
rm(list=ls())
set.seed(18041119)

### Set the sample size for the four cells in the two-way balanced design
n1 <- 32
n2 <- 32
n3 <- 32
n4 <- 32

### Initializing the four cells in factor 1, factor 2 and the dependent variable
f1_1 <- f1_2 <- f1_3 <- f1_4 <- NULL
f2_1 <- f2_2 <- f2_3 <- f2_4 <- NULL
dv_1 <- dv_2 <- dv_3 <- dv_4 <- NULL

### Set the values for the number of cells, the conditions of true null hypothesis and false
nulls, ### the standard deviation, and number of replications for each pattern.
num_cells <- 4
num_patterns <- 4
m_pattern_0 <- c(0.0,0.0,0.0,0.0)
m_pattern_1 <- c(0.0,0.5,0.0,0.5)
m_pattern_2 <- c(0.5,1.0,0.0,0.5)
m_pattern_3 <- c(0.0,0.0,1.0,0.0)
sd_pattern <- c(1.0,1.0,1.0,1.0)
n_rep <- 20,000

### Set the elements in the array, there are 3 conditions of the nonnull effects, 5 MHT
methods, ### and 3 types of error rates
max_pat <- 3
max_meth <- 5
max_types <- 3
errors <- array (NA,dim=c(max_pat,max_meth,max_types))

### Initializing the p values in 5 MHT procedures, each has 3 conditions and is a matrix
### of n_rep*3
p_unadjust <- NULL
p_unadjust$p1 <- matrix(0,n_rep,3)
p_unadjust$p2 <- matrix(0,n_rep,3)
p_unadjust$p3 <- matrix(0,n_rep,3)
p_unadjust$p4 <- matrix(0,n_rep,3)
p_bonf <- p_holm <- p_hochberg <- p_BH <- p_unadjust
```
### Set all rates for the first three patterns using MHT procedures

#### Unadjusted alpha
#### Initializing each pattern
```r
sampfwer.un.p1 <- 0
sampfwer.un.p2 <- 0
sampfwer.un.p3 <- 0
bhfwer.un.p1 <- 0
bhfwer.un.p2 <- 0
bhfwer.un.p3 <- 0
```

#### # set sampfwer, bhfwer, fdr, and totalrejections for unadjusted method for pattern 1
```r
sampfwer.rate.un.p1 <- 0
bhfwer.rate.un.p1 <- 0
fdr.un.p1 <- 0
totalrejections.un.p1 <- 0
```

#### # set sampfwer, bhfwer, fdr, and totalrejections for unadjusted method for pattern 2
```r/nsampfwer.rate.un.p2 <- 0
  bhfwer.rate.un.p2 <- 0
  fdr.un.p2 <- 0
  totalrejections.un.p2 <- 0
```

#### # set sampfwer, bhfwer, fdr, and totalrejections for unadjusted method for pattern 3
```r
sampfwer.rate.un.p3 <- 0
  bhfwer.rate.un.p3 <- 0
  fdr.un.p3 <- 0
  totalrejections.un.p3 <- 0
```

#### Bonferroni Procedure
#### Initializing each pattern
```r
sampfwer.bonf.p1 <- 0
sampfwer.bonf.p2 <- 0
sampfwer.bonf.p3 <- 0
bhfwer.bonf.p1 <- 0
bhfwer.bonf.p2 <- 0
bhfwer.bonf.p3 <- 0
```

#### #set sampfwer, bhfwer, fdr, and totalrejections for pattern 1 in the Bonferroni procedure
```r
sampfwer.rate.bonf.p1 <- 0
  bhfwer.rate.bonf.p1 <- 0
  fdr.bonf.p1 <- 0
  totalrejections.bonf.p1 <- 0
```
#set sampfwer, bhfwer, fdr, and totalrejections for pattern 2 in the Bonferroni procedure
sampfwer.rate.bonf.p2 <- 0
bhfwer.rate.bonf.p2 <- 0
fdr.bonf.p2 <- 0
totalrejections.bonf.p2 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 3 in the Bonferroni procedure
sampfwer.rate.bonf.p3 <- 0
bhfwer.rate.bonf.p3 <- 0
fdr.bonf.p3 <- 0
totalrejections.bonf.p3 <- 0

### Holm's procedure
#Initializing each pattern
sampfwer.holm.p1 <- 0
sampfwer.holm.p2 <- 0
sampfwer.holm.p3 <- 0
bhfwer.holm.p1 <- 0
bhfwer.holm.p2 <- 0
bhfwer.holm.p3 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 1 in the Holm procedure
sampfwer.rate.holm.p1 <- 0
bhfwer.rate.holm.p1 <- 0
fdr.holm.p1 <- 0
totalrejections.holm.p1 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 2 in the Holm procedure
sampfwer.rate.holm.p2 <- 0
bhfwer.rate.holm.p2 <- 0
fdr.holm.p2 <- 0
totalrejections.holm.p2 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 3 in the Holm procedure
sampfwer.rate.holm.p3 <- 0
bhfwer.rate.holm.p3 <- 0
fdr.holm.p3 <- 0
totalrejections.holm.p3 <- 0

### Hochberg procedure
#Initializing each pattern
sampfwer.hochberg.p1 <- 0
sampfwer.hochberg.p2 <- 0
sampfwer.hochberg.p3 <- 0

### Hochberg procedure
#Initializing each pattern
sampfwer.hochberg.p1 <- 0
sampfwer.hochberg.p2 <- 0
sampfwer.hochberg.p3 <- 0
bhfwer.hochberg.p1 <- 0
bhfwer.hochberg.p2 <- 0
bhfwer.hochberg.p3 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 1 in the Hochberg procedure
sampfwer.rate.hochberg.p1 <- 0
bhfwer.rate.hochberg.p1 <- 0
fdr.hochberg.p1 <- 0
totalrejections.hochberg.p1 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 2 in the Hochberg procedure
sampfwer.rate.hochberg.p2 <- 0
bhfwer.rate.hochberg.p2 <- 0
fdr.hochberg.p2 <- 0
totalrejections.hochberg.p2 <- 0
#set sampfwer, bhfwer, fdr, and totalrejections for pattern 3 in the Hochberg procedure
sampfwer.rate.hochberg.p3 <- 0
bhfwer.rate.hochberg.p3 <- 0
fdr.hochberg.p3 <- 0
totalrejections.hochberg.p3 <- 0

###B-H Procedure
#Initializing each pattern
sampfwer.BH.p1 <- 0
sampfwer.BH.p2 <- 0
sampfwer.BH.p3 <- 0
bhfwer.BH.p1 <- 0
bhfwer.BH.p2 <- 0
bhfwer.BH.p3 <- 0

#set sampfwer, bhfwer, fdr, and totalrejections for pattern 1 in the B-H procedure
sampfwer.rate.BH.p1 <- 0
bhfwer.rate.BH.p1 <- 0
fdr.BH.p1 <- 0
totalrejections.BH.p1 <- 0
#set sampfwer, bhfwer, fdr, and totalrejections for pattern 2 in the B-H procedure
sampfwer.rate.BH.p2 <- 0
bhfwer.rate.BH.p2 <- 0
fdr.BH.p2 <- 0
totalrejections.BH.p2 <- 0
#set sampfwer, bhfwer, fdr, and totalrejections for pattern 3 in the B-H procedure
sampfwer.rate.BH.p3 <- 0
bhfwer.rate.BH.p3 <- 0
fdr.BH.p3 <- 0
totalrejections.BH.p3 <- 0

### set all power rates for all last three patterns using MHT procedures
### unadjusted alpha
swpower.un.p2 <- 0  # set swpower for unadjusted pattern 2
swpower.un.p3 <- 0  # set swpower for unadjusted pattern 3
swpower.un.p4 <- 0  # set swpower for unadjusted pattern 4

twpower.un.p2 <- 0  # set twpower for unadjusted pattern 2
swpower.un.p3 <- 0  # set twpower for unadjusted pattern 3
swpower.un.p4 <- 0  # set twpower for unadjusted pattern 4
swpower.rate.un.p2 <- 0  # set swpower rate for unadjusted pattern 2
twpower.rate.un.p2 <- 0  # set swpower rate for unadjusted pattern 2
tdr.un.p2 <- 0  # set tdr for unadjusted pattern 2
p_totalrejections.un.p2 <- 0  # set p_totalrejections unadjusted pattern 2
swpower.rate.un.p3 <- 0
swpower.rate.un.p4 <- 0
twpower.rate.un.p3 <- 0
swpower.rate.un.p4 <- 0
swpower.rate.un.p3 <- 0
swpower.rate.un.p4 <- 0
twpower.rate.un.p3 <- 0
swpower.rate.un.p4 <- 0
twpower.rate.un.p4 <- 0

### Bonferroni Procedure
swpower.bonf.p2 <- 0  # set swpower for bonferroni pattern 2
swpower.bonf.p3 <- 0  # set swpower for bonferroni pattern 3
swpower.bonf.p4 <- 0  # set swpower for bonferroni pattern 4

twpower.bonf.p2 <- 0  # set twpower for bonferroni pattern 2
twpower.bonf.p3 <- 0  # set twpower for bonferroni pattern 3
twpower.bonf.p4 <- 0  # set twpower for bonferroni pattern 4

swpower.rate.bonf.p2 <- 0
twpower.rate.bonf.p2 <- 0
tdr.bonf.p2 <- 0
p_totalrejections.bonf.p2 <- 0
swpower.rate.bonf.p3 <- 0
twpower.rate.bonf.p3 <- 0
tdr.bonf.p3 <- 0
p_totalrejections.bonf.p3 <- 0
swpower.rate.bonf.p4 <- 0
twpower.rate.bonf.p4 <- 0
tdr.bonf.p4 <- 0
p_totalrejections.bonf.p4 <- 0

#### Holm's procedure
swpower.holm.p2 <- 0  # set swpower for Holm pattern 2
swpower.holm.p3 <- 0  # set swpower for Holm pattern 3
swpower.holm.p4 <- 0  # set swpower for Holm pattern 4
twpower.holm.p2 <- 0  # set twpower for Holm pattern 2
twpower.holm.p3 <- 0  # set twpower for Holm pattern 3
twpower.holm.p4 <- 0  # set twpower for Holm pattern 4
swpower.rate.holm.p2 <- 0
swpower.rate.holm.p3 <- 0
swpower.rate.holm.p4 <- 0
twpower.rate.holm.p2 <- 0
twpower.rate.holm.p3 <- 0
twpower.rate.holm.p4 <- 0
tdr.holm.p2 <- 0
p_totalrejections.holm.p2 <- 0

swpower.holm.p3 <- 0
swpower.holm.p4 <- 0
tdr.holm.p3 <- 0
tdr.holm.p4 <- 0
p_totalrejections.holm.p3 <- 0
p_totalrejections.holm.p4 <- 0

#### Hochberg procedure
swpower.hochberg.p2 <- 0  # set swpower for hochberg pattern 2
swpower.hochberg.p3 <- 0  # set swpower for hochberg pattern 3
swpower.hochberg.p4 <- 0  # set swpower for hochberg pattern 4
twpower.hochberg.p2 <- 0  # set twpower for hochberg pattern 2
twpower.hochberg.p3 <- 0  # set twpower for hochberg pattern 3
twpower.hochberg.p4 <- 0  # set twpower for hochberg pattern 4
swpower.rate.hochberg.p2 <- 0
swpower.rate.hochberg.p2 <- 0
tdr.hochberg.p2 <- 0
p_totalrejections.hochberg.p2 <- 0

swpower.rate.hochberg.p3 <- 0
twpower.rate.hochberg.p3 <- 0
tdr.hochberg.p3 <- 0
p_totalrejections.hochberg.p3 <- 0

swpower.rate.hochberg.p4 <- 0
twpower.rate.hochberg.p4 <- 0
tdr.hochberg.p4 <- 0
p_totalrejections.hochberg.p4 <- 0

###BH Procedure
swpower.BH.p2 <- 0  # set swpower for BH pattern 2
swpower.BH.p3 <- 0  # set swpower for BH pattern 3
swpower.BH.p4 <- 0  # set swpower for BH pattern 4

twpower.BH.p2 <- 0  # set twpower for BH pattern 2
 twpower.BH.p3 <- 0  # set twpower for BH pattern 3
 twpower.BH.p4 <- 0  # set twpower for BH pattern 4

swpower.rate.BH.p2 <- 0
twpower.rate.BH.p2 <- 0
tdr.BH.p2 <- 0
p_totalrejections.BH.p2 <- 0

swpower.rate.BH.p3 <- 0
twpower.rate.BH.p3 <- 0
tdr.BH.p3 <- 0
p_totalrejections.BH.p3 <- 0

swpower.rate.BH.p4 <- 0
twpower.rate.BH.p4 <- 0
tdr.BH.p4 <- 0
p_totalrejections.BH.p4 <- 0

### The following sets up the cell mean, standard deviation, cell size min, and max for 4 patterns
cm1 <- matrix(0,n_rep,4)
cm2 <- matrix(0,n_rep,4)
cm3 <- matrix(0,n_rep,4)
cm4 <- matrix(0,n_rep,4)
sd1 <- matrix(0,n_rep,4)
sd2 <- matrix(0,n_rep,4)
sd3 <- matrix(0,n_rep,4)
sd4 <- matrix(0,n_rep,4)
ss1 <- matrix(0,n_rep,4)
ss2 <- matrix(0,n_rep,4)
ss3 <- matrix(0,n_rep,4)
ss4 <- matrix(0,n_rep,4)
min1 <- matrix(0,n_rep,4)
min2 <- matrix(0,n_rep,4)
min3 <- matrix(0,n_rep,4)
min4 <- matrix(0,n_rep,4)
max1 <- matrix(0,n_rep,4)
max2 <- matrix(0,n_rep,4)
max3 <- matrix(0,n_rep,4)
max4 <- matrix(0,n_rep,4)
###Loop for each pattern
for (ixpat in 1: num_patterns)
{
    ##find the pattern and paste the corresponding values
    if (ixpat==1) {mean_pattern <- m_pattern_0}
    else if (ixpat==2) {mean_pattern <- m_pattern_1}
    else if (ixpat==3) {mean_pattern <- m_pattern_2}
    else if (ixpat==4) {mean_pattern <- m_pattern_3}

    m1 <- mean_pattern[1]; m2 <- mean_pattern[2]; m3 <- mean_pattern[3]; m4 <-
    mean_pattern[4]
    s1 <- sd_pattern[1];       s2 <- sd_pattern[2];        s3 <- sd_pattern[3];        s4 <-
    sd_pattern[4]

    for (i in 1:num_cells)
    {
        if (i==1) {f1_1 <- rep (1,each=n1,times=1)
                 f2_1 <- rep (1,each=n1,times=1)}
        if (i==2) {f1_2 <- rep (1,each=n2,times=1)
                 f2_2 <- rep (2,each=n2,times=1)}
        if (i==3) {f1_3 <- rep (2,each=n3,times=1)
                 f2_3 <- rep (1,each=n3,times=1)}
        if (i==4) {f1_4 <- rep (2,each=n4,times=1)
                 f2_4 <- rep (2,each=n4,times=1)}
    } # end cell

    f1 <- c(f1_1,f1_2,f1_3,f1_4)         #combine four cell values into f1
f2 <- c(f2_1,f2_2,f2_3,f2_4)  #combine four cell values into f2
ff1 <- as.factor(f1)          #regard f1 as factor ff1
ff2 <- as.factor(f2)          #regard f2 as factor ff2

###Create some temporary variables to record the results
temp_unadjust <- matrix(0,n_rep,3)
temp_bonf <- temp_holm <- temp_hochberg <- temp_BH <- temp_unadjust

##loop for each repetition
for (tt in 1:n_rep)
{
  ##loop for each cell
  for (i in 1:num_cells)
  {
    if (i==1) dv_1 <- rnorm(n1,m1,s1)
    if (i==2) dv_2 <- rnorm(n2,m2,s2)
    if (i==3) dv_3 <- rnorm(n3,m3,s3)
    if (i==4) dv_4 <- rnorm(n4,m4,s4)
  } # end cell

dv <- c(dv_1,dv_2,dv_3,dv_4)  #combine four cell dv into one dv
dataset <- data.frame(dv,f1,f2)  #combine dv, f1,f2 into one dataset

###Calculate cell means, standard deviation, sample size per cell, min, and max for each pattern
if (ixpat==1) {cm1[tt,1] <- mean (dv_1)
cm1[tt,2] <- mean (dv_2)
cm1[tt,3] <- mean (dv_3)
cm1[tt,4] <- mean (dv_4)
sd1[tt,1] <- sd (dv_1)
sd1[tt,2] <- sd (dv_2)
sd1[tt,3] <- sd (dv_3)
sd1[tt,4] <- sd (dv_4)
ss1[tt,1] <- length(dv_1)
ss1[tt,2] <- length(dv_2)
ss1[tt,3] <- length(dv_3)
ss1[tt,4] <- length(dv_4)
min1[tt,1] <- min (dv_1)
min1[tt,2] <- min (dv_2)
min1[tt,3] <- min (dv_3)
min1[tt,4] <- min (dv_4)
max1[tt,1] <- max (dv_1)
max1[tt,2] <- max (dv_2)
max1[tt,3] <- max (dv_3)
max1[tt,4] <- max (dv_4)
else if (ixpat==2)  
{cm2[tt,1] <- mean (dv_1)
cm2[tt,2] <- mean (dv_2)
cm2[tt,3] <- mean (dv_3)
cm2[tt,4] <- mean (dv_4)
sd2[tt,1] <- sd (dv_1)
sd2[tt,2] <- sd (dv_2)
sd2[tt,3] <- sd (dv_3)
sd2[tt,4] <- sd (dv_4)
ss2[tt,1] <- length(dv_1)
ss2[tt,2] <- length(dv_2)
ss2[tt,3] <- length(dv_3)
ss2[tt,4] <- length(dv_4)
min2[tt,1] <- min (dv_1)
min2[tt,2] <- min (dv_2)
min2[tt,3] <- min (dv_3)
min2[tt,4] <- min (dv_4)
max2[tt,1] <- max (dv_1)
max2[tt,2] <- max (dv_2)
max2[tt,3] <- max (dv_3)
max2[tt,4] <- max (dv_4)
}
else if (ixpat==3)  
{cm3[tt,1] <- mean (dv_1)
cm3[tt,2] <- mean (dv_2)
cm3[tt,3] <- mean (dv_3)
cm3[tt,4] <- mean (dv_4)
sd3[tt,1] <- sd (dv_1)
sd3[tt,2] <- sd (dv_2)
sd3[tt,3] <- sd (dv_3)
sd3[tt,4] <- sd (dv_4)
ss3[tt,1] <- length(dv_1)
ss3[tt,2] <- length(dv_2)
ss3[tt,3] <- length(dv_3)
ss3[tt,4] <- length(dv_4)
min3[tt,1] <- min (dv_1)
min3[tt,2] <- min (dv_2)
min3[tt,3] <- min (dv_3)
min3[tt,4] <- min (dv_4)
max3[tt,1] <- max (dv_1)
max3[tt,2] <- max (dv_2)
max3[tt,3] <- max (dv_3)
max3[tt,4] <- max (dv_4)
}
else if (ixpat==4)  
{cm4[tt,1] <- mean (dv_1)
cm4[tt,2] <- mean (dv_2)
cm4[tt,3] <- mean (dv_3)
cm4[tt,4] <- mean (dv_4)
sd4[tt,1] <- sd (dv_1)
sd4[tt,2] <- sd (dv_2)
sd4[tt,3] <- sd (dv_3)
sd4[tt,4] <- sd (dv_4)
ss4[tt,1] <- length(dv_1)
ss4[tt,2] <- length(dv_2)
ss4[tt,3] <- length(dv_3)
ss4[tt,4] <- length(dv_4)
min4[tt,1] <- min (dv_1)
min4[tt,2] <- min (dv_2)
min4[tt,3] <- min (dv_3)
min4[tt,4] <- min (dv_4)
max4[tt,1] <- max (dv_1)
max4[tt,2] <- max (dv_2)
max4[tt,3] <- max (dv_3)
max4[tt,4] <- max (dv_4)

mod <- lm(dv~ff1*ff2,data=dataset)                              #conduct linear model in R
anova <- anova(mod)                                                     #conduct anova analysis in R
pvalues <- anova$Pr[1:3]                                               #pull out p values
###Conduct MHT in R
    temp_unadjust[tt,] <- round(p.adjust(pvalues, "none"), 5)
    temp_bonf[tt,] <- round(p.adjust(pvalues, "bonferroni"), 5)
    temp_holm[tt,] <- round(p.adjust(pvalues, "holm"), 5)
    temp_hochberg[tt,] <- round(p.adjust(pvalues, "hochberg"), 5)
    temp_BH[tt,] <- round(p.adjust(pvalues, "BH"), 5)

}# end replication

#####count false rejections to calculate the Type I error rates

if  (ixpat==1)
{
    p_unadjust$p1 <- temp_unadjust
    p_bonf$p1 <- temp_bonf
    p_holm$p1 <- temp_holm
    p_hochberg$p1 <- temp_hochberg
    p_BH$p1 <- temp_BH

    for (itest in 1:n_rep)
    {
        ## count sampfwer in the unadjusted method for pattern 1
        if        (p_unadjust$p1[itest,1] < 0.05)    {sampfwer.un.p1 <- sampfwer.un.p1 + 1}
else if (p_unadjust$p1[itest,2] < 0.05) {sampfwer.un.p1 <- sampfwer.un.p1 + 1}
else if (p_unadjust$p1[itest,3] < 0.05) {sampfwer.un.p1 <- sampfwer.un.p1 + 1}
else {sampfwer.un.p1 <- sampfwer.un.p1}
}

{ ## count bhfwer in the unadjusted method for pattern 1
if (p_unadjust$p1[itest,1] < 0.05) {bhfwer.un.p1 <- bhfwer.un.p1 + 1}
if (p_unadjust$p1[itest,2] < 0.05) {bhfwer.un.p1 <- bhfwer.un.p1 + 1}
if (p_unadjust$p1[itest,3] < 0.05) {bhfwer.un.p1 <- bhfwer.un.p1 + 1}
}

{ ## count sampfwer in the bonferroni method for pattern 1
if   (p_bonf$p1[itest,1] < 0.05)      {sampfwer.bonf.p1 <- sampfwer.bonf.p1 + 1}
else if (p_bonf$p1[itest,2] < 0.05) {sampfwer.bonf.p1 <- sampfwer.bonf.p1 + 1}
else if (p_bonf$p1[itest,3] < 0.05) {sampfwer.bonf.p1 <- sampfwer.bonf.p1 + 1}
else                                                      {sampfwer.bonf.p1 <- sampfwer.bonf.p1}
}

{ ## count bhfwer in the bonferroni method for pattern 1
if (p_bonf$p1[itest,1] < 0.05)           {bhfwer.bonf.p1 <- bhfwer.bonf.p1 + 1}
if (p_bonf$p1[itest,2] < 0.05)           {bhfwer.bonf.p1 <- bhfwer.bonf.p1 + 1}
if (p_bonf$p1[itest,3] < 0.05)           {bhfwer.bonf.p1 <- bhfwer.bonf.p1 + 1}
}

{ ## count sampfwer in the holm method for pattern 1
if   (p_holm$p1[itest,1] < 0.05)      {sampfwer.holm.p1 <- sampfwer.holm.p1 + 1}
else if (p_holm$p1[itest,2] < 0.05) {sampfwer.holm.p1 <- sampfwer.holm.p1 + 1}
else if (p_holm$p1[itest,3] < 0.05)  {sampfwer.holm.p1 <- sampfwer.holm.p1 + 1}
else                                                   {sampfwer.holm.p1 <- sampfwer.holm.p1}
}

{ ## count bhfwer in the holm method for pattern 1
if (p_holm$p1[itest,1] < 0.05)           {bhfwer.holm.p1 <- bhfwer.holm.p1 + 1}
if (p_holm$p1[itest,2] < 0.05)           {bhfwer.holm.p1 <- bhfwer.holm.p1 + 1}
if (p_holm$p1[itest,3] < 0.05)           {bhfwer.holm.p1 <- bhfwer.holm.p1 + 1}
}

{ ## count sampfwer in the hochberg method for pattern 1
if  (p_hochberg$p1[itest,1] < 0.05)  {sampfwer.hochberg.p1 <- sampfwer.hochberg.p1 + 1}
else if (p_hochberg$p1[itest,2] < 0.05{sampfwer.hochberg.p1 <- sampfwer.hochberg.p1 + 1}
else if (p_hochberg$p1[itest,3] < 0.05){sampfwer.hochberg.p1 <- sampfwer.hochberg.p1 + 1}
else                                                       {sampfwer.hochberg.p1 <- sampfwer.hochberg.p1}
}

{ ## count bhfwer in the hochberg method for pattern 1
if (p_hochberg$p1[itest,1] < 0.05)        {bhfwer.hochberg.p1 <- bhfwer.hochberg.p1 + 1}
if (p_hochberg$p1[itest,2] < 0.05)        {bhfwer.hochberg.p1 <-
bhfwer.hochberg.p1 + 1}
if (p_hochberg$p1[itest,3] < 0.05)        {bhfwer.hochberg.p1 <-
bhfwer.hochberg.p1 + 1}
}
} ### count sampfwer in the BH method for pattern 1
if       (p_BH$p1[itest,1] < 0.05)        {sampfwer.BH.p1 <- sampfwer.BH.p1 + 1}
else if (p_BH$p1[itest,2] < 0.05)        {sampfwer.BH.p1 <- sampfwer.BH.p1 + 1}
else if (p_BH$p1[itest,3] < 0.05)        {sampfwer.BH.p1 <- sampfwer.BH.p1 + 1}
else                                                      {sampfwer.BH.p1 <- sampfwer.BH.p1}
}
} ### count bhfwer in the BH method for pattern 1
if (p_BH$p1[itest,1] < 0.05)            {bhfwer.BH.p1 <- bhfwer.BH.p1 + 1}
if (p_BH$p1[itest,2] < 0.05)            {bhfwer.BH.p1 <- bhfwer.BH.p1 + 1}
if (p_BH$p1[itest,3] < 0.05)            {bhfwer.BH.p1 <- bhfwer.BH.p1 + 1}
}
}
}

else if (ixpat==2)
{
    p_unadjust$p2 <- temp_unadjust
    p_bonf$p2 <- temp_bonf
    p_holm$p2 <- temp_holm
    p_hochberg$p2 <- temp_hochberg
    p_BH$p2 <- temp_BH

for (itest in 1:n_rep)
{
    
    if       (p_unadjust$p2[itest,1] < 0.05) {sampfwer.un.p2 <- sampfwer.un.p2 + 1}
    else if (p_unadjust$p2[itest,3] < 0.05) {sampfwer.un.p2 <- sampfwer.un.p2 + 1}
    else                                                  {sampfwer.un.p2 <- sampfwer.un.p2}

    
    if (p_unadjust$p2[itest,1] < 0.05)     {bhfwer.un.p2 <- bhfwer.un.p2 + 1}
    if (p_unadjust$p2[itest,3] < 0.05)     {bhfwer.un.p2 <- bhfwer.un.p2 + 1}

    
    if     (p_bonf$p2[itest,1] < 0.05)        {sampfwer.bonf.p2 <- sampfwer.bonf.p2 + 1}
    else if(p_bonf$p2[itest,3] < 0.05)     {sampfwer.bonf.p2 <- sampfwer.bonf.p2 + 1}
    else                                                    {sampfwer.bonf.p2 <- sampfwer.bonf.p2}

}
if (p_bonf$p2[itest,1] < 0.05) {bhfwer.bonf.p2 <- bhfwer.bonf.p2 + 1}
if (p_bonf$p2[itest,3] < 0.05) {bhfwer.bonf.p2 <- bhfwer.bonf.p2 + 1}

if (p_holm$p2[itest,1] < 0.05) {sampfwer.holm.p2 <- sampfwer.holm.p2 + 1}
else if (p_holm$p2[itest,3] < 0.05) {sampfwer.holm.p2 <- sampfwer.holm.p2 + 1}
else {sampfwer.holm.p2 <- sampfwer.holm.p2}

if (p_hochberg$p2[itest,1] < 0.05) {sampfwer.hochberg.p2 <- sampfwer.hochberg.p2 + 1}
else if (p_hochberg$p2[itest,3] < 0.05) {sampfwer.hochberg.p2 <- sampfwer.hochberg.p2 + 1}
else {sampfwer.hochberg.p2 <- sampfwer.hochberg.p2}

if (p_BH$p2[itest,1] < 0.05) {sampfwer.BH.p2 <- sampfwer.BH.p2 + 1}
else if (p_BH$p2[itest,3] < 0.05) {sampfwer.BH.p2 <- sampfwer.BH.p2 + 1}
else {sampfwer.BH.p2 <- sampfwer.BH.p2}

else {
p_unadjust$p3 <- temp_unadjust
p_bonf$p3 <- temp_bonf
p_holm$p3 <- temp_holm
p_hochberg$p3 <- temp_hochberg
}
p_BH$p3 <- temp_BH

for (itest in 1:n_rep)
{
    if (p_unadjust$p3[itest,3] < 0.05) {sampfwer.un.p3 <- sampfwer.un.p3 + 1}
    if (p_unadjust$p3[itest,3] < 0.05) {bhfwer.un.p3 <- bhfwer.un.p3 + 1}

    if (p_bonf$p3[itest,3] < 0.05) {sampfwer.bonf.p3 <- sampfwer.bonf.p3 + 1}
    if (p_bonf$p3[itest,3] < 0.05) {bhfwer.bonf.p3 <- bhfwer.bonf.p3 + 1}

    if (p_holm$p3[itest,3] < 0.05) {sampfwer.holm.p3 <- sampfwer.holm.p3 + 1}
    if (p_holm$p3[itest,3] < 0.05) {bhfwer.holm.p3 <- bhfwer.holm.p3 + 1}

    if (p_hochberg$p3[itest,3] < 0.05) {sampfwer.hochberg.p3 <- sampfwer.hochberg.p3 + 1}
    if (p_hochberg$p3[itest,3] < 0.05) {bhfwer.hochberg.p3 <- bhfwer.hochberg.p3 + 1}

    if (p_BH$p3[itest,3] < 0.05) {sampfwer.BH.p3 <- sampfwer.BH.p3 + 1}
    if (p_BH$p3[itest,3] < 0.05) {bhfwer.BH.p3 <- bhfwer.BH.p3 + 1}
}

####count true rejections to calculate power rate

if (ixpat==2)
{
    p_unadjust$p2 <- temp_unadjust
    p_bonf$p2 <- temp_bonf
    p_holm$p2 <- temp_holm
    p_hochberg$p2 <- temp_hochberg
    p_BH$p2 <- temp_BH

    for (itest in 1:n_rep)
    {
        if (p_unadjust$p2[itest,2] < 0.05) {swpower.un.p2 <- swpower.un.p2 + 1}
        if (p_unadjust$p2[itest,2] < 0.05) {twpower.un.p2 <- twpower.un.p2 + 1}
    }
if (p_bonf$p2[itest,2] < 0.05)           {swpower.bonf.p2 <- swpower.bonf.p2 + 1}
if (p_bonf$p2[itest,2] < 0.05)           {twpower.bonf.p2 <- twpower.bonf.p2 + 1}

if (p_holm$p2[itest,2] < 0.05)           {swpower.holm.p2 <- swpower.holm.p2 + 1}
if (p_holm$p2[itest,2] < 0.05)           {twpower.holm.p2 <- twpower.holm.p2 + 1}

if (p_hochberg$p2[itest,2] < 0.05)       {swpower.hochberg.p2 <- swpower.hochberg.p2 + 1}
if (p_hochberg$p2[itest,2] < 0.05)       {twpower.hochberg.p2 <- twpower.hochberg.p2 + 1}

if (p_BH$p2[itest,2] < 0.05)              {swpower.BH.p2 <- swpower.BH.p2 + 1}
if (p_BH$p2[itest,2] < 0.05)              {twpower.BH.p2 <- twpower.BH.p2 + 1}

else if (ixpat==3)
  {
    p_unadjust$p3 <- temp_unadjust
    p_bonf$p3 <- temp_bonf
    p_holm$p3 <- temp_holm
    p_hochberg$p3 <- temp_hochberg
    p_BH$p3 <- temp_BH

    for (itest in 1:n_rep)
      {
        if     (p_unadjust$p3[itest,1] < 0.05) {swpower.un.p3 <- swpower.un.p3 + 1}
        else if(p_unadjust$p3[itest,2] < 0.05) {swpower.un.p3 <- swpower.un.p3 + 1}
        else                                   {swpower.un.p3 <- swpower.un.p3}
      }
      
      if (p_unadjust$p3[itest,1] < 0.05)     {twpower.un.p3 <- twpower.un.p3 + 1}
      if (p_unadjust$p3[itest,2] < 0.05)     {twpower.un.p3 <- twpower.un.p3 + 1}
      
      if (p_hochberg$p3[itest,2] < 0.05)     {swpower.hochberg.p3 <- swpower.hochberg.p3 + 1}
      if (p_hochberg$p3[itest,2] < 0.05)     {twpower.hochberg.p3 <- twpower.hochberg.p3 + 1}
      
      if (p_BH$p3[itest,2] < 0.05)            {swpower.BH.p3 <- swpower.BH.p3 + 1}
      if (p_BH$p3[itest,2] < 0.05)            {twpower.BH.p3 <- twpower.BH.p3 + 1}
  }
if (p_bonf$p3[itest,1] < 0.05) {swpower.bonf.p3 <- swpower.bonf.p3 + 1}
else if(p_bonf$p3[itest,2] < 0.05) {swpower.bonf.p3 <- swpower.bonf.p3 + 1}
else
  {swpower.bonf.p3 <- swpower.bonf.p3}
}

if (p_bonf$p3[itest,1] < 0.05) {twpower.bonf.p3 <- twpower.bonf.p3 + 1}
if (p_bonf$p3[itest,2] < 0.05) {twpower.bonf.p3 <- twpower.bonf.p3 + 1}

if (p_holm$p3[itest,1] < 0.05) {swpower.holm.p3 <- swpower.holm.p3 + 1}
else if(p_holm$p3[itest,2] < 0.05) {swpower.holm.p3 <- swpower.holm.p3 + 1}
else
  {swpower.holm.p3 <- swpower.holm.p3}

if (p_holm$p3[itest,1] < 0.05) {twpower.holm.p3 <- twpower.holm.p3 + 1}
if (p_holm$p3[itest,2] < 0.05) {twpower.holm.p3 <- twpower.holm.p3 + 1}

if (p_hochberg$p3[itest,1] < 0.05) {swpower.hochberg.p3 <- swpower.hochberg.p3 + 1}
else if(p_hochberg$p3[itest,2] < 0.05) {swpower.hochberg.p3 <- swpower.hochberg.p3 + 1}
else
  {swpower.hochberg.p3 <- swpower.hochberg.p3}

if (p_hochberg$p3[itest,1] < 0.05) {twpower.hochberg.p3 <- twpower.hochberg.p3 + 1}
if (p_hochberg$p3[itest,2] < 0.05) {twpower.hochberg.p3 <- twpower.hochberg.p3 + 1}

if (p_BH$p3[itest,1] < 0.05) {swpower.BH.p3 <- swpower.BH.p3 + 1}
else if(p_BH$p3[itest,2] < 0.05) {swpower.BH.p3 <- swpower.BH.p3 + 1}
else
  {swpower.BH.p3 <- swpower.BH.p3}

if (p_BH$p3[itest,1] < 0.05) {twpower.BH.p3 <- twpower.BH.p3 + 1}
if (p_BH$p3[itest,2] < 0.05) {twpower.BH.p3 <- twpower.BH.p3 + 1}
else if (ixpat==4)
{
    p_unadjust$p4 <- temp_unadjust
    p_bonf$p4 <- temp_bonf
    p_holm$p4 <- temp_holm
    p_hochberg$p4 <- temp_hochberg
    p_BH$p4 <- temp_BH
}
for (itest in 1:n_rep)
{
    if      (p_unadjust$p4[itest,1] < 0.05)    {swpower.un.p4 <- swpower.un.p4 + 1}
    else if (p_unadjust$p4[itest,2] < 0.05)  {swpower.un.p4 <- swpower.un.p4 + 1}
    else if (p_unadjust$p4[itest,3] < 0.05)  {swpower.un.p4 <- swpower.un.p4 + 1}
    else                                                        {swpower.un.p4 <- swpower.un.p4}
}
    if (p_unadjust$p4[itest,1] < 0.05)    {twpower.un.p4 <- twpower.un.p4 + 1}
    if (p_unadjust$p4[itest,2] < 0.05)  {twpower.un.p4 <- twpower.un.p4 + 1}
    if (p_unadjust$p4[itest,3] < 0.05)  {twpower.un.p4 <- twpower.un.p4 + 1}
}
    if      (p_bonf$p4[itest,1] < 0.05)        {swpower.bonf.p4 <- swpower.bonf.p4 + 1}
    else if (p_bonf$p4[itest,2] < 0.05)      {swpower.bonf.p4 <- swpower.bonf.p4 + 1}
    else if (p_bonf$p4[itest,3] < 0.05)      {swpower.bonf.p4 <- swpower.bonf.p4 + 1}
    else                                                      {swpower.bonf.p4 <- swpower.bonf.p4}
}
    if (p_bonf$p4[itest,1] < 0.05)           {twpower.bonf.p4 <- twpower.bonf.p4 + 1}
    if (p_bonf$p4[itest,2] < 0.05)           {twpower.bonf.p4 <- twpower.bonf.p4 + 1}
    if (p_bonf$p4[itest,3] < 0.05)           {twpower.bonf.p4 <- twpower.bonf.p4 + 1}
}
    if      (p_holm$p4[itest,1] < 0.05)      {swpower.holm.p4 <- swpower.holm.p4 + 1}
    else if (p_holm$p4[itest,2] < 0.05)   {swpower.holm.p4 <- swpower.holm.p4 + 1}
    else if (p_holm$p4[itest,3] < 0.05)   {swpower.holm.p4 <- swpower.holm.p4 + 1}
    else                                                   {swpower.holm.p4 <- swpower.holm.p4}
}
    if (p_holm$p4[itest,1] < 0.05)           {twpower.holm.p4 <- twpower.holm.p4 + 1}
    if (p_holm$p4[itest,2] < 0.05)           {twpower.holm.p4 <- twpower.holm.p4 + 1}
    if (p_holm$p4[itest,3] < 0.05)           {twpower.holm.p4 <- twpower.holm.p4 + 1}
if (p_hochberg$p4[itest,1] < 0.05) {swpower.hochberg.p4 <- swpower.hochberg.p4 + 1}
else if (p_hochberg$p4[itest,2] < 0.05) {swpower.hochberg.p4 <- swpower.hochberg.p4 + 1}
else if (p_hochberg$p4[itest,3] < 0.05) {swpower.hochberg.p4 <- swpower.hochberg.p4 + 1}
else {swpower.hochberg.p4 <- swpower.hochberg.p4}

if (p_hochberg$p4[itest,1] < 0.05) {twpower.hochberg.p4 <- twpower.hochberg.p4 + 1}
else if (p_hochberg$p4[itest,2] < 0.05) {twpower.hochberg.p4 <- twpower.hochberg.p4 + 1}
else if (p_hochberg$p4[itest,3] < 0.05) {twpower.hochberg.p4 <- twpower.hochberg.p4 + 1}
else {twpower.hochberg.p4 <- twpower.hochberg.p4 + 1}

if (p_BH$p4[itest,1] < 0.05) {swpower.BH.p4 <- swpower.BH.p4 + 1}
else if (p_BH$p4[itest,2] < 0.05) {swpower.BH.p4 <- swpower.BH.p4 + 1}
else if (p_BH$p4[itest,3] < 0.05) {swpower.BH.p4 <- swpower.BH.p4 + 1}
else {swpower.BH.p4 <- swpower.BH.p4}

if (p_BH$p4[itest,1] < 0.05) {twpower.BH.p4 <- twpower.BH.p4 + 1}
else if (p_BH$p4[itest,2] < 0.05) {twpower.BH.p4 <- twpower.BH.p4 + 1}
else if (p_BH$p4[itest,3] < 0.05) {twpower.BH.p4 <- twpower.BH.p4 + 1}
else {twpower.BH.p4 <- twpower.BH.p4 + 1}

### Save the data for each pattern

if (ixpat==1)
  {write.table (dataset,file="mydatabalance21new.csv",quote=FALSE,sep=" ",na=" ",row.names=TRUE,col.names=TRUE)}
else if (ixpat==2)
  {write.table (dataset,file="mydatabalance22new.csv",quote=FALSE,sep=" ",na=" ",row.names=TRUE,col.names=TRUE)}
else if (ixpat==3)
{write.table (dataset,file="mydatabalance23new.csv",quote=FALSE,sep="",na="",row.names=TRUE,col.names=TRUE)}

else if (ixpat==4)
{write.table (dataset,file="mydatabalance24new.csv",quote=FALSE,sep="",na="",row.names=TRUE,col.names=TRUE)}

} # end pattern loop

####Count and show sampfwer, bhfwer with all patterns in MHT
##pattern 1
sampfwer.un.p1
bhfwer.un.p1
p_unadjust$p1
sampfwer.bonf.p1
bhfwer.bonf.p1
p_bonf$p1
sampfwer.holm.p1
bhfwer.holm.p1
p_holm$p1
sampfwer.hochberg.p1
bhfwer.hochberg.p1
p_hochberg$p1
sampfwer.BH.p1
bhfwer.BH.p1
p_BH$p1

##pattern 2
sampfwer.un.p2
bhfwer.un.p2
p_unadjust$p2
sampfwer.bonf.p2
bhfwer.bonf.p2
p_bonf$p2
sampfwer.holm.p2
bhfwer.holm.p2
p_holm$p2
sampfwer.hochberg.p2
bhfwer.hochberg.p2
p_hochberg$p2
sampfwer.BH.p2
bhfwer.BH.p2
p_BH$p2
### Calculate proportion for each pattern with sampfwer, bhfwer, and fdr

#### pattern 3

```r
sampfwer.un.p3
bhfwer.un.p3
p_unadjust$p3	sampfwer.bonf.p3
bhfwer.bonf.p3
p_bonf$p3	sampfwer.holm.p3
bhfwer.holm.p3
p_holm$p3	sampfwer.hochberg.p3
bhfwer.hochberg.p3
p_hochberg$p3	sampfwer.BH.p3
bhfwer.BH.p3
p_BHi$p3
```

#### Calculate proportion for each pattern with sampfwer, bhfwer, and fdr

#### pattern 1

```r
sampfwer.rate.un.p1
bhfwer.rate.un.p1 <- bhfwer.un.p1/(n_rep*3)
bhfwer.rate.un.p1
totalrejections.un.p1 <- sum(p_unadjust$p1[, 1:3]< 0.05)
totalrejections.un.p1
fdr.un.p1
ers[1,1,1] <- sampfwer.rate.un.p1
ers[1,1,2] <- bhfwer.rate.un.p1
ers[1,1,3] <- fdr.un.p1
sampfwer.rate.bonf.p1 <- sampfwer.bonf.p1/n_rep
sampfwer.rate.bonf.p1
bhfwer.rate.bonf.p1 <- bhfwer.bonf.p1/(n_rep*3)
bhfwer.rate.bonf.p1
totalrejections.bonf.p1 <- sum(p_bonf$p1[, 1:3]< 0.05)
totalrejections.bonf.p1
fdr.bonf.p1 <- bhfwer.bonf.p1/totalrejections.bonf.p1
fdr.bonf.p1
ers[1,2,1] <- sampfwer.rate.bonf.p1
ers[1,2,2] <- bhfwer.rate.bonf.p1
ers[1,2,3] <- fdr.bonf.p1
sampfwer.rate.holm.p1 <- sampfwer.holm.p1/n_rep
sampfwer.rate.holm.p1
```
bhfwer.rate.holm.p1 <- bhfwer.holm.p1/(n_rep*3)
bhfwer.rate.holm.p1
totalrejections.holm.p1 <- sum(p_holm$p1[, 1:3]< 0.05)
totalrejections.holm.p1
fdr.holm.p1 <- bhfwer.holm.p1/totalrejections.holm.p1
fdr.holm.p1
errors[1,3,1] <- sampfwer.rate.holm.p1
errors[1,3,2] <- bhfwer.rate.holm.p1
errors[1,3,3] <- fdr.holm.p1
sampfwer.rate.hochberg.p1 <- sampfwer.hochberg.p1/n_rep
sampfwer.rate.hochberg.p1
bhfwer.rate.hochberg.p1 <- bhfwer.hochberg.p1/(n_rep*3)
bhfwer.rate.hochberg.p1
totalrejections.hochberg.p1 <- sum(p_hochberg$p1[, 1:3]< 0.05)
totalrejections.hochberg.p1
fdr.hochberg.p1 <- bhfwer.hochberg.p1/totalrejections.hochberg.p1
fdr.hochberg.p1
errors[1,4,1] <- sampfwer.rate.hochberg.p1
errors[1,4,2] <- bhfwer.rate.hochberg.p1
errors[1,4,3] <- fdr.hochberg.p1
sampfwer.rate.BH.p1 <- sampfwer.BH.p1/n_rep
sampfwer.rate.BH.p1
bhfwer.rate.BH.p1 <- bhfwer.BH.p1/(n_rep*3)
bhfwer.rate.BH.p1
totalrejections.BH.p1 <- sum(p_BH$p1[, 1:3]< 0.05)
totalrejections.BH.p1
fdr.BH.p1 <- bhfwer.BH.p1/totalrejections.BH.p1
fdr.BH.p1
errors[1,5,1] <- sampfwer.rate.BH.p1
errors[1,5,2] <- bhfwer.rate.BH.p1
errors[1,5,3] <- fdr.BH.p1
##pattern2
sampfwer.rate.un.p2 <- sampfwer.un.p2/n_rep
sampfwer.rate.un.p2
bhfwer.rate.un.p2 <- bhfwer.un.p2/(n_rep*2)
bhfwer.rate.un.p2
totalrejections.un.p2 <- sum(p_unadjust$p2[, 1:3]< 0.05)
totalrejections.un.p2
fdr.un.p2
errors[2,1,1] <- sampfwer.rate.un.p2
errors[2,1,2] <- bhfwer.rate.un.p2
errors[2,1,3] <- fdr.un.p2
sampfwer.rate.bonf.p2 <- sampfwer.bonf.p2/n_rep
sampfwer.rate.bonf.p2
bhfwer.rate.bonf.p2 <- bhfwer.bonf.p2/(n_rep*2)
bhfwer.rate.bonf.p2
totalrejections.bonf.p2 <- sum(p_bonf$p2[, 1:3]< 0.05)
totalrejections.bonf.p2
fdr.bonf.p2 <- bhfwer.bonf.p2/totalrejections.bonf.p2
fdr.bonf.p2
errors[2,2,1] <- sampfwer.rate.bonf.p2
errors[2,2,2] <- bhfwer.rate.bonf.p2
errors[2,2,3] <- fdr.bonf.p2
sampfwer.rate.holm.p2 <- sampfwer.holm.p2/n_rep
sampfwer.rate.holm.p2
bhfwer.rate.holm.p2 <- bhfwer.holm.p2/(n_rep*2)
bhfwer.rate.holm.p2
totalrejections.holm.p2 <- sum(p_holm$p2[, 1:3]< 0.05)
totalrejections.holm.p2
fdr.holm.p2 <- bhfwer.holm.p2/totalrejections.holm.p2
fdr.holm.p2
errors[2,3,1] <- sampfwer.rate.holm.p2
errors[2,3,2] <- bhfwer.rate.holm.p2
errors[2,3,3] <- fdr.holm.p2
sampfwer.rate.hochberg.p2 <- sampfwer.hochberg.p2/n_rep
sampfwer.rate.hochberg.p2
bhfwer.rate.hochberg.p2 <- bhfwer.hochberg.p2/(n_rep*2)
bhfwer.rate.hochberg.p2
totalrejections.hochberg.p2 <- sum(p_hochberg$p2[, 1:3]< 0.05)
totalrejections.hochberg.p2
fdr.hochberg.p2 <- bhfwer.hochberg.p2/totalrejections.hochberg.p2
fdr.hochberg.p2
errors[2,4,1] <- sampfwer.rate.hochberg.p2
errors[2,4,2] <- bhfwer.rate.hochberg.p2
errors[2,4,3] <- fdr.hochberg.p2
sampfwer.rate.BH.p2 <- sampfwer.BH.p2/n_rep
sampfwer.rate.BH.p2
bhfwer.rate.BH.p2 <- bhfwer.BH.p2/(n_rep*2)
bhfwer.rate.BH.p2
totalrejections.BH.p2 <- sum(p_BH$p2[, 1:3]< 0.05)
totalrejections.BH.p2
fdr.BH.p2 <- bhfwer.BH.p2/totalrejections.BH.p2
fdr.BH.p2
errors[2,5,1] <- sampfwer.rate.BH.p2
errors[2,5,2] <- bhfwer.rate.BH.p2
errors[2,5,3] <- fdr.BH.p2
## pattern 3
sampfwer.rate.un.p3
bhfwer.rate.un.p3 <- bhfwer.un.p3/(n_rep)
bhfwer.rate.un.p3
totalrejections.un.p3 <- sum(p_unadjust$p3[, 1:3]< 0.05)
totalrejections.un.p3
fdr.un.p3
errors[3,1,1] <- sampfwer.rate.un.p3
errors[3,1,2] <- bhfwer.rate.un.p3
errors[3,1,3] <- fdr.un.p3
sampfwer.rate.bonf.p3 <- sampfwer.bonf.p3/n_rep
sampfwer.rate.bonf.p3
bhfwer.rate.bonf.p3 <- bhfwer.bonf.p3/(n_rep)
bhfwer.rate.bonf.p3
totalrejections.bonf.p3 <- sum(p_bonf$p3[, 1:3]< 0.05)
totalrejections.bonf.p3
fdr.bonf.p3
errors[3,2,1] <- sampfwer.rate.bonf.p3
errors[3,2,2] <- bhfwer.rate.bonf.p3
errors[3,2,3] <- fdr.bonf.p3
sampfwer.rate.holm.p3 <- sampfwer.holm.p3/n_rep
sampfwer.rate.holm.p3
bhfwer.rate.holm.p3 <- bhfwer.holm.p3/(n_rep)
bhfwer.rate.holm.p3
totalrejections.holm.p3 <- sum(p_holm$p3[, 1:3]< 0.05)
totalrejections.holm.p3
fdr.holm.p3
errors[3,3,1] <- sampfwer.rate.holm.p3
errors[3,3,2] <- bhfwer.rate.holm.p3
errors[3,3,3] <- fdr.holm.p3
sampfwer.rate.hochberg.p3 <- sampfwer.hochberg.p3/n_rep
sampfwer.rate.hochberg.p3
bhfwer.rate.hochberg.p3 <- bhfwer.hochberg.p3/(n_rep)
bhfwer.rate.hochberg.p3
totalrejections.hochberg.p3 <- sum(p_hochberg$p3[, 1:3]< 0.05)
totalrejections.hochberg.p3
fdr.hochberg.p3
errors[3,4,1] <- sampfwer.rate.hochberg.p3
errors[3,4,2] <- bhfwer.rate.hochberg.p3
errors[3,4,3] <- fdr.hochberg.p3
sampfwer.rate.BH.p3 <- sampfwer.BH.p3/n_rep
sampfwer.rate.BH.p3
bhfwer.rate.BH.p3 <- bhfwer.BH.p3/(n_rep)
bhfwer.rate.BH.p3
totalrejections.BH.p3 <- sum(p_BH$p3[, 1:3]< 0.05)
totalrejections.BH.p3
fdr.BH.p3 <- bhfwer.BH.p3/totalrejections.BH.p3
fdr.BH.p3
errors[3,5,1] <- sampfwer.rate.BH.p3
errors[3,5,2] <- bhfwer.rate.BH.p3
errors[3,5,3] <- fdr.BH.p3

##### Count true rejections for the last three patterns

swpower.un.p2
twpower.un.p2
#p_unadjust$p2
swpower.bonf.p2
twpower.bonf.p2
#p_bonf$p2
swpower.holm.p2
twpower.holm.p2
#p_holm$p2
swpower.hochberg.p2
twpower.hochberg.p2
#p_hochberg$p2
swpower.BH.p2
twpower.BH.p2
#p_BH$p2

swpower.un.p3
twpower.un.p3
#p_unadjust$p3
swpower.bonf.p3
twpower.bonf.p3
#p_bonf$p3
swpower.holm.p3
twpower.holm.p3
#p_holm$p3
swpower.hochberg.p3
twpower.hochberg.p3
#p_hochberg$p3
```r
## calculate power rate for each pattern with swpower, twpower, and tdr

## pattern2

swpower.rate.un.p2 <- swpower.un.p2/n_rep
swpower.rate.un.p2
twpower.rate.un.p2 <- twpower.un.p2/(n_rep*1)
twpower.rate.un.p2

p_totalrejections.un.p2 <- sum(p_unadjust$p2[, 1:3]< 0.05)
p_totalrejections.un.p2
tdr.un.p2

swpower.rate.bonf.p2 <- swpower.bonf.p2/n_rep
swpower.rate.bonf.p2
twpower.rate.bonf.p2 <- twpower.bonf.p2/(n_rep*1)
twpower.rate.bonf.p2

p_totalrejections.bonf.p2 <- sum(p_bonf$p2[, 1:3]< 0.05)
p_totalrejections.bonf.p2
tdr.bonf.p2 <- twpower.bonf.p2/p_totalrejections.bonf.p2
tdr.bonf.p2
```
swpower.rate.holm.p2 <- swpower.holm.p2/n_rep
swpower.rate.holm.p2
twpower.rate.holm.p2 <- twpower.holm.p2/(n_rep*1)
twpower.rate.holm.p2

p_totalrejections.holm.p2 <- sum(p_holm$p2[,1:3]< 0.05)
p_totalrejections.holm.p2
tdr.holm.p2 <- twpower.holm.p2/p_totalrejections.holm.p2
tdr.holm.p2

swpower.rate.hochberg.p2 <- swpower.hochberg.p2/n_rep
swpower.rate.hochberg.p2
twpower.rate.hochberg.p2 <- twpower.hochberg.p2/(n_rep*1)
twpower.rate.hochberg.p2

p_totalrejections.hochberg.p2 <- sum(p_hochberg$p2[,1:3]< 0.05)
p_totalrejections.hochberg.p2
tdr.hochberg.p2 <- twpower.hochberg.p2/p_totalrejections.hochberg.p2
tdr.hochberg.p2

swpower.rate.BH.p2 <- swpower.BH.p2/n_rep
swpower.rate.BH.p2
twpower.rate.BH.p2 <- twpower.BH.p2/(n_rep*1)
twpower-rate.BH.p2

p_totalrejections.BH.p2 <- sum(p_BH$p2[,1:3]< 0.05)
p_totalrejections.BH.p2
tdr.BH.p2 <- twpower.BH.p2/p_totalrejections.BH.p2
tdr.BH.p2

###pattern 3
swpower.rate.un.p3
twpower-rate.un.p3<- twpower.un.p3/(n_rep*2)
twpower-rate.un.p3

p_totalrejections.un.p3 <- sum(p_unadjust$p3[,1:3]< 0.05)
p_totalrejections.un.p3
tdr.un.p3

swpower.rate.bonf.p3 <- swpower.bonf.p3/n_rep
swpower.rate.bonf.p3

twpower.rate.bonf.p3 <- twpower.bonf.p3/(n_rep*2)
twpower.rate.bonf.p3

p_totalrejections.bonf.p3 <- sum(p_bonf$p3[, 1:3]< 0.05)
p_totalrejections.bonf.p3
tdr.bonf.p3

swpower.rate.holm.p3 <- swpower.holm.p3/n_rep
swpower.rate.holm.p3

twpower.rate.holm.p3 <- twpower.holm.p3/(n_rep*2)
twpower.rate.holm.p3

p_totalrejections.holm.p3 <- sum(p_holm$p3[, 1:3]< 0.05)
p_totalrejections.holm.p3
tdr.holm.p3

swpower.rate.hochberg.p3 <- swpower.hochberg.p3/n_rep
swpower.rate.hochberg.p3

twpower.rate.hochberg.p3 <- twpower.hochberg.p3/(n_rep*2)
twpower.rate.hochberg.p3

p_totalrejections.hochberg.p3 <- sum(p_hochberg$p3[, 1:3]< 0.05)
p_totalrejections.hochberg.p3
tdr.hochberg.p3

swpower.rate.BH.p3 <- swpower.BH.p3/n_rep
swpower.rate.BH.p3

twpower.rate.BH.p3 <- twpower.BH.p3/(n_rep*2)
twpower.rate.BH.p3

p_totalrejections.BH.p3 <- sum(p_BH$p3[, 1:3]< 0.05)
p_totalrejections.BH.p3
tdr.BH.p3 <- twpower.BH.p3/p_totalrejections.BH.p3
tdr.BH.p3

## pattern4

swpower.rate.un.p4 <- swpower.un.p4/n_rep
swpower.rate.un.p4
twpower.rate.un.p4 <- twpower.un.p4/(n_rep*3)
twpower.rate.un.p4

p_totalrejections.un.p4 <- sum(p_unadjust$p4[, 1:3]< 0.05)
p_totalrejections.un.p4
tdr.un.p4

swpower.rate.bonf.p4 <- swpower.bonf.p4/n_rep
swpower.rate.bonf.p4
twpower.rate.bonf.p4 <- twpower.bonf.p4/(n_rep*3)
twpower.rate.bonf.p4

p_totalrejections.bonf.p4 <- sum(p_bonf$p4[, 1:3]< 0.05)
p_totalrejections.bonf.p4
tdr.bonf.p4 <- twpower.bonf.p4/p_totalrejections.bonf.p4
tdr.bonf.p4

swpower.rate.holm.p4 <- swpower.holm.p4/n_rep
swpower.rate.holm.p4
twpower.rate.holm.p4 <- twpower.holm.p4/(n_rep*3)
twpower.rate.holm.p4

p_totalrejections.holm.p4 <- sum(p_holm$p4[, 1:3]< 0.05)
p_totalrejections.holm.p4
tdr.holm.p4 <- twpower.holm.p4/p_totalrejections.holm.p4
tdr.holm.p4

swpower.rate.hochberg.p4 <- swpower.hochberg.p4/n_rep
swpower.rate.hochberg.p4
twpower.rate.hochberg.p4 <- twpower.hochberg.p4/(n_rep*3)
twpower.rate.hochberg.p4

p_totalrejections.hochberg.p4 <- sum(p_hochberg$p4[, 1:3]< 0.05)
p_totalrejections.hochberg.p4
tdr.hochberg.p4 <- twpower.hochberg.p4/p_totalrejections.hochberg.p4
tdr.hochberg.p4

swpower.rate.BH.p4 <- swpower.BH.p4/n_rep
swpower.rate.BH.p4
twpower.rate.BH.p4 <- twpower.BH.p4/(n_rep*3)
twpower.rate.BH.p4

p_totalrejections.BH.p4 <- sum(p_BH$p4[, 1:3] < 0.05)
p_totalrejections.BH.p4
tdr.BH.p4 <- twpower.BH.p4/p_totalrejections.BH.p4
tdr.BH.p4

save.image("2waycode_1220_balance_complete.Rdata")
## Pattern 1

```r
mydatabalance21 <- read.csv("mydatabalance21new.csv")
cellmean <- tapply(mydatabalance21$dv, data.frame(f1, f2), mean, na.rm = TRUE)
cellmean

cm1[100,]
cellsd <- tapply(mydatabalance21$dv, data.frame(f1, f2), sd, na.rm = TRUE)
cellsd

sd1[100,]
table(f1, f2)

ss1[100,]
cellmin <- tapply(mydatabalance21$dv, data.frame(f1, f2), min, na.rm = TRUE)
cellmin

min1[100,]
cellmax <- tapply(mydatabalance21$dv, data.frame(f1, f2), max, na.rm = TRUE)
cellmax

max1[100,]
mod <- lm(dv ~ f1 * f2, data = mydatabalance21)
mod

anova <- anova(mod)
anova

pvalues <- anova$Pr[1:3]
pvalues

p_unadjust$p1[100,]
```

## Pattern 2

```r
mydatabalance22 <- read.csv("mydatabalance22new.csv")
cellmean <- tapply(mydatabalance22$dv, data.frame(f1, f2), mean, na.rm = TRUE)
cellmean

cm2[100,]
cellsd <- tapply(mydatabalance22$dv, data.frame(f1, f2), sd, na.rm = TRUE)
cellsd

sd2[100,]
table(f1, f2)

ss2[100,]
cellmin <- tapply(mydatabalance22$dv, data.frame(f1, f2), min, na.rm = TRUE)
cellmin

min2[100,]
cellmax <- tapply(mydatabalance22$dv, data.frame(f1, f2), max, na.rm = TRUE)
cellmax

max2[100,]
mod <- lm(dv ~ f1 * f2, data = mydatabalance22)
```
mod
anova <- anova(mod)
anova
pvalues <- anova$Pr[1:3]
pvalues
p_unadjust$p2[100,]

##Pattern3
mydatabalance23<-read.csv("mydatabalance23new.csv")
cellmean <- tapply(mydatabalance23$dv,data.frame(f1,f2),mean,na.rm=TRUE)
cellmean
cm3[100,]
cellsd <- tapply(mydatabalance23$dv,data.frame(f1,f2),sd,na.rm=TRUE)
cellsd
csd3[100,]
table(f1,f2)
ss3[100,]
cellmin <- tapply(mydatabalance23$dv,data.frame(f1,f2),min,na.rm=TRUE)
cellmin
min3[100,]
cellmax <- tapply(mydatabalance23$dv,data.frame(f1,f2),max,na.rm=TRUE)
cellmax
max3[100,]
mod <- lm(dv~f1*f2,data=mydatabalance23)
mod
anova <- anova(mod)
anova
pvalues <- anova$Pr[1:3]
pvalues
p_unadjust$p3[100,]

##Use the following code after loading the workspace
sum(p_unadjust$p1[,1]<0.05)
sum(p_unadjust$p1[,2]<0.05)
sum(p_unadjust$p1[,3]<0.05)
sampfwer.un.p1
bhfwer.un.p1
sum(p_unadjust$p2[,1]<0.05)
sum(p_unadjust$p2[,2]<0.05)
sum(p_unadjust$p2[,3]<0.05)
sampfwer.un.p2
bhfwer.un.p2
sum(p_unadjust$p3[,1]<0.05)
sum(p_unadjust$p3[,2]<0.05)
sum(p_unadjust$p3[,3]<0.05)
sampfwer.un.p3
bhfwer.un.p3
sum(p_bonf$p1[,1]<0.05)
sum(p_bonf$p1[,2]<0.05)
sum(p_bonf$p1[,3]<0.05)
sampfwer.bonf.p1
bhfwer.bonf.p1
sum(p_bonf$p2[,1]<0.05)
sum(p_bonf$p2[,2]<0.05)
sum(p_bonf$p2[,3]<0.05)
sampfwer.bonf.p2
bhfwer.bonf.p2
sum(p_bonf$p3[,1]<0.05)
sum(p_bonf$p3[,2]<0.05)
sum(p_bonf$p3[,3]<0.05)
sampfwer.bonf.p3
bhfwer.bonf.p3
sum(p_holm$p1[,1]<0.05)
sum(p_holm$p1[,2]<0.05)
sum(p_holm$p1[,3]<0.05)
sampfwer.holm.p1
bhfwer.holm.p1
sum(p_holm$p2[,1]<0.05)
sum(p_holm$p2[,2]<0.05)
sum(p_holm$p2[,3]<0.05)
sampfwer.holm.p2
bhfwer.holm.p2
sum(p_holm$p3[,1]<0.05)
sum(p_holm$p3[,2]<0.05)
sum(p_holm$p3[,3]<0.05)
sampfwer.holm.p3
bhfwer.holm.p3
sum(p_hochberg$p1[,1]<0.05)
sum(p_hochberg$p1[,2]<0.05)
sum(p_hochberg$p1[,3]<0.05)
sampfwer.hochberg.p1
bhfwer.hochberg.p1
sum(p_hochberg$p2[,1]<0.05)
sum(p_hochberg$p2[,2]<0.05)
sum(p_hochberg$p2[,3]<0.05)
sampfwer.hochberg.p2
bhfwer.hochberg.p2
sum(p_hochberg$p3[,1]<0.05)
sum(p_hochberg$p3[,2]<0.05)
sampfwer.hochberg.p3
bhfwer.hochberg.p3
sum(p_BH$p1[,1]<0.05)
sampfwer.BH.p1
bhfwer.BH.p1
sum(p_BH$p2[,1]<0.05)
sampfwer.BH.p2
bhfwer.BH.p2
sum(p_BH$p3[,1]<0.05)
sampfwer.BH.p3
bhfwer.BH.p3
sum(p_unadjust$p1[,1]<0.05)/n_rep
sum(p_unadjust$p1[,2]<0.05)/n_rep
sum(p_unadjust$p1[,3]<0.05)/n_rep
sum(p_unadjust$p2[,1]<0.05)/n_rep
sum(p_unadjust$p2[,2]<0.05)/n_rep
sum(p_unadjust$p2[,3]<0.05)/n_rep
sum(p_unadjust$p3[,1]<0.05)/n_rep
sum(p_unadjust$p3[,2]<0.05)/n_rep
sum(p_unadjust$p3[,3]<0.05)/n_rep
sum(p_bonf$p1[,1]<0.05)/n_rep
sum(p_bonf$p1[,2]<0.05)/n_rep
sum(p_bonf$p1[,3]<0.05)/n_rep
sum(p_bonf$p2[,1]<0.05)/n_rep
sum(p_bonf$p2[,2]<0.05)/n_rep
sum(p_bonf$p2[,3]<0.05)/n_rep
sum(p_bonf$p3[,1]<0.05)/n_rep
sum(p_bonf$p3[,2]<0.05)/n_rep
sum(p_bonf$p3[,3]<0.05)/n_rep
sum(p_holm$p1[,1]<0.05)/n_rep
sum(p_holm$p1[,2]<0.05)/n_rep
sum(p_holm$p1[,3]<0.05)/n_rep
sum(p_holm$p2[,1]<0.05)/n_rep
sum(p_holm$p2[,2]<0.05)/n_rep
sum(p_holm$p2[,3]<0.05)/n_rep
sum(p_holm$p3[,1]<0.05)/n_rep
sum(p_holm$p3[,2]<0.05)/n_rep
sum(p_holm$p3[,3]<0.05)/n_rep
sum(p_hochberg$p1[,1]<0.05)/n_rep
sum(p_hochberg$p1[,2]<0.05)/n_rep
sum(p_hochberg$p1[,3]<0.05)/n_rep
sum(p_hochberg$p2[,1]<0.05)/n_rep
sum(p_hochberg$p2[,2]<0.05)/n_rep
sum(p_hochberg$p2[,3]<0.05)/n_rep
sum(p_hochberg$p3[,1]<0.05)/n_rep
sum(p_hochberg$p3[,2]<0.05)/n_rep
sum(p_hochberg$p3[,3]<0.05)/n_rep
sum(p_BH$p1[,1]<0.05)/n_rep
sum(p_BH$p1[,2]<0.05)/n_rep
sum(p_BH$p1[,3]<0.05)/n_rep
sum(p_BH$p2[,1]<0.05)/n_rep
sum(p_BH$p2[,2]<0.05)/n_rep
sum(p_BH$p2[,3]<0.05)/n_rep
sum(p_BH$p3[,1]<0.05)/n_rep
sum(p_BH$p3[,2]<0.05)/n_rep
sum(p_BH$p3[,3]<0.05)/n_rep