Type I Error Rates and Power of Multiple Hypothesis Testing Procedures in Factorial ANOVA

Qian An **Deyu Xu** **Gordon P. Brooks**
Ohio University

There are numerous General Linear Model (GLM) statistical designs that may require multiple hypothesis testing (MHT) procedures that control the Type I error inflation that occurs with multiple tests. This study investigated familywise error rates (FWER) and statistical power rates of several alpha-adjustment MHT procedures in factorial ANOVA, but results apply broadly to GLM procedures. Of four MHT procedures investigated, the Hochberg procedure performed most efficiently in terms of Type I error and power, slightly better than Holm. The Holm procedure, however, may be the better choice because of less restrictive assumptions. FWER concerns were raised with the Benjamini-Hochberg procedure.

In educational research, many statistical analyses are conducted using null hypothesis significance tests. Often, only a single null 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 an intervention and a control. To answer these research questions, various types of *t* tests could be adopted under a given level of significance, or α. In statistics, if the null hypothesis is incorrectly rejected, Type I error occurs. That is, there is a 5% of chance that researchers can make a Type I error for a single hypothesis test when α = .05.

However, when *k* multiple hypotheses are tested, the *k* separate null hypothesis significance tests are often performed each at α = .05. The probability of making at least one Type I error when multiple, independent true null hypotheses are tested is defined as $p=1-(1-\alpha)^k$, where α is the nominal Type I error rate (often .05) and *k* is the number of independent hypothesis tests performed (Hochberg & Tamhane, 1987; Maxwell & Delaney, 2000; Schochet, 2008; Stevens, 2002; 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). Therefore, the Type I error rate is inflated across multiple hypothesis tests (MHT).

Many researchers are familiar with post hoc multiple comparison procedures, which are able to maintain the familywise Type I error rate at α when a set of post hoc comparisons is made among sample means, following a significant one-way ANOVA. These procedures (e.g., Tukey; Scheffé) adjust nominal alpha for each test so that the probability that at least one of the significant null hypotheses is a Type I error remains at a given familywise alpha.

However, multiple comparison procedures (MCPs) tested in one-way ANOVA represent only one area that uses MHT. In fact, MHT can be conducted in several other areas, such as (a) multiple tests of correlations among variables, (b) univariate and multivariate post hoc tests in MANOVA methods, (c) multiple chi-square tests in differential item functioning, (d) tests of multiple coefficients in multiple regression, (e) repeated measures post hoc tests, and (f) tests of multiple sources of variation in factorial ANOVA. Several kinds of general purpose MHT procedures have been made available for researchers, including the Bonferroni procedure, the Holm (1979) procedure, the Hochberg (1988) procedure, and the Benjamini-Hochberg (1995) procedure. Like MCP methods, the use of general MHT procedures is to control the Type I error rate when testing multiple hypotheses simultaneously.

Several scholars have 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 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). 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, Type I error rates and statistical power rates were investigated to evaluate several MHT procedures in this study.

Type I Error Rates

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 when working with MCPs. PTER is the probability that any one of the hypothesis tests is falsely rejected (Hinkle, Wiersma, & Jurs, 2003; Ryan, 1959), whereas FWER is the probability that at least one statistical null hypothesis is incorrectly rejected in a given family (Hochberg & Tamhane; Ryan, 1959; Shaffer; Toothaker). Part of the complexity of handling FWER is due to the uncertain and varying definitions of "family."

Ryan (1959) and Miller (1981) argued that the family of hypotheses includes all those tested on the results from a single experiment. Maxwell and Delaney (2000) proposed that each main effect and interaction in a factorial ANOVA should constitute its own family. However, several researchers also suggested treating all tests in the factorial designs as one family (Games, 1971; Ryan, 1959; Stevens, 1999). 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).

Relationship to Statistical Power. Statistical power is the probability to detect a significant effect when a hypothesis test is indeed false in the population. We know from Cohen (1988) and others that Type I error, effect size, sample size, and Type II error (and therefore statistical power) are functionally related. For example, given (a) an alpha level, (b) an expected effect size, (c) a sample size, then (d) statistical power can be determined for the analysis.

False Discovery Rate

Another type of error rate, the False Discovery Rate (FDR), was introduced in Benjamini and Hochberg (1995). Figure 1 is adapted from Benjamini and Hochberg, who illustrated the FDR. In Figure 1, m_0 is the number of true null hypotheses, $m-m_0$ is the number of false null hypotheses, *and* m is the total number of hypothesis tests. U is the number of true correct acceptances, V is the number of false rejections (i.e., Type I errors), T is the number of false acceptances (i.e., Type II errors), and S is the number of correct rejections. The total number of acceptances is $m-R$, and R is total rejections. When all hypothesis tests are true null, $m = m_0$.

		Sample-based decision		
		Accepted	Rejected	Total
Population condition	True Null	U	V	m_0
	Non-True Null	T	S	$m-m_0$
	Total	$m-R$	R	m

Figure 1. Definition of Errors

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 of as the expected proportion of the false rejections over the number of true null hypotheses. It can be written as:

$$FWER = E(\frac{V}{m_0} \mid m_0 > 0)$$

FDR is the proportion of false rejections of true null hypotheses from among the total rejections on all hypotheses, whether true or false in the population. Therefore, it is expressed as:

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

Benjamini and Hochberg (1995) described FDR as the expected proportion of false rejections among all rejections. FDR may be desirable because, for example, a few errors of inference should be tolerable in exploratory research; thus, one can use the less stringent FDR method of control. Also, Shaffer (1995) suggested that "on the average, only a proportion α of the rejected hypotheses are true ones" (p. 567). FDR gives researchers another way to think about rejections.

When all null hypotheses are true (i.e., "completely null" scenario), FDR equals FWER and therefore “control of the FDR implies control of the FWER in the weak sense” (Benjamini & Hochberg, 1995, p. 291). Whenever the number of true null (m_0) is smaller than the number of total hypotheses (m), the FDR is smaller than or equal to the FWER. That is, controlling FDR also controls FWER. FDR is more powerful than FWER in many circumstances (Benjamini & Hochberg; Keselman, Cribbie, & Holland, 2002; Schochet, 2008; William, Jones, & Tukey, 1999).

Whether to Adjust Alpha

There is a lack of consensus among scholars about whether to adjust the alpha level in social and behavioral science. Some researchers argue that alpha should never be adjusted. 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)

Indeed, without adequate definition of family, researchers can adjust alpha ad absurdum (e.g., lifetime alpha, doling out portions of alpha to users of national databases).

However, many researchers support adjusting alpha in research some or all of the time (Games, 1971; Keselman et al., 2002; Ryan, 1959, 1960; Westfall & Young, 1993). Keppel (1991) suggested that "familywise error is the inevitable penalty associated with conducting additional comparisons" (p. 247). 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). Some guidelines have been suggested. For example, What Works Clearinghouse (2008), the National Center for Education Statistics (2009), the National Assessment of Educational Progress, the Institute of Education Sciences, and others recommend adjusting alpha for multiple hypothesis tests (Schochet, 2008; Williams et al., 1999).

Procedure in Multiple Hypothesis Testing

Hancock and Klockars (1996) indicated that "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 statistical power. Some MHT procedures only control the Type I error rate when all null hypotheses are true in the population (i.e., the "completely null" case); other procedures work well also 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).

A three-way ANOVA example is adapted from Neter, Kutner, Nachtsheim, and Wasserman (1996) to help explain the techniques studied here (results reported in Table 1 and Table 2). 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*). As these examples are explored, it should be noted that one can perform adjusted tests equivalently by comparing obtained significance values to adjusted alpha values or by comparing adjusted *p* values with nominal alpha.

Unadjusted Alpha. Traditionally, hypothesis testing is often conducted at a specified nominal alpha level for each test. Each *p* value is compared to the nominal alpha level in this method. For example, if .05 is set, all *p* values produced in one statistical analysis should be compared to .05 and final conclusions made based on those comparisons. However, Type I error rate inflation occurs when a set of hypotheses is tested simultaneously if each hypothesis test is compared to .05. The probability of making at least one Type I error when multiple and independent null hypotheses are true is $p=1-(1-\alpha)^k$, where overall nominal α is often .05 and *k* is the number of hypothesis tests (Hochberg & Tamhane, 1987; Maxwell & Delaney, 2000; Schochet, 2008; Stevens, 2002; Toothaker, 1993). 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.

Bonferroni. In the Bonferroni procedure, the nominal familywise alpha (α) level is divided by the number of hypothesis tests. Using this approach results in the actual FWER remaining below nominal familywise alpha. Consequently, each hypothesis is tested at the same alpha level $\alpha_i = \alpha/n$, where *n* is the number of null hypothesis significance tests performed. The null hypothesis is rejected if the individual *p* value is less than α/n---or equivalently, if np is less than α then the null hypothesis is rejected. For the Neter et al. (1996) example above, each null hypothesis is rejected if the obtained $7p$ is less than .05. As a result, only the main effects *A* and *B* are significant, because their adjusted *p* values are less than the

Table 1. Three-Way ANOVA Output

Source	Sum of Squares	df	Mean Square	*F* Ratio	*p* values
A	168.584	1	168.584	18.059	0.001
B	242.570	1	242.570	25.984	0.000
C	74.384	1	74.384	7.968	0.012
AB	13.650	1	13.650	1.462	0.244
AC	11.070	1	11.070	1.186	0.292
BC	76.454	1	76.454	8.190	0.011
ABC	1.870	1	1.870	0.200	0.660
Error	149.367	16	9.335		

Table 2. Using the Bonferroni, Holm, Hochberg, and Benjamini-Hochberg Procedures

Hypothesis	Sorted unadjusted *p* value	Bonferroni-adjusted *p* value	Holm-adjusted *p* value	Hochberg-adjusted *p* value	Benjamini-Hochberg-adjusted *p* value
B	0.000*	0.000*	0.000*	0.000*	0.000*
A	0.001*	0.007*	0.006*	0.006*	0.004*
BC	0.011*	0.077	0.055	0.055*	0.026*
C	0.012*	0.084	0.048	0.048*	0.021*
AB	0.244	1.000	0.732	0.732	0.342
AC	0.292	1.000	0.584	0.584	0.341
ABC	0.660	1.000	0.660	0.660	0.660

Note. * indicates significant test *adjusted p* < .05 (values of 1.000 represent *adjusted p* values ≥ 1.0).

familywise alpha of .05. That is, Gender (Factor A) had unadjusted $p = .001$, so $7p = .007$, which is less than $\alpha = .05$.

The Bonferroni correction has become the standard and most well-known approach for controlling the FWER in the strong sense (Cai, 2006; Shaffer, 1995; Schochet, 2008). 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). It is a simple but general procedure that does not require any constraining assumptions (Hochberg & Benjamini, 1990), but is conservative because the adjusted alpha level becomes very small if there are many hypothesis tests (Field, 2005; Games, 1971; Hancock & Klockars, 1996; Holm, 1979). Therefore, the Bonferroni procedure suffers from lower power in many situations (e.g., Hochberg & Benjamini; Schochet).

A number of modifications have been made to the original Bonferroni procedure. In particular, a number of sequential strategies have been developed that improve power slightly compared to the traditional Bonferroni, but at the cost of complexity (Hancock & Klockars, 1996; Shaffer, 1995). For example, sequential methods are not able to have simple confidence intervals like simultaneous procedures can (Hancock & Klockars; Holland & Copenhaver, 1987; Toothaker, 1993). We focus on two common approaches: Holm (1979) and Hochberg (1988).

Holm. Holm (1979) introduced a sequentially rejective Bonferroni procedure. Like the Bonferroni procedure, the Holm procedure has no constraining assumptions and controls the Type I error rate in a strong sense (Holland & Copenhaver, 1987; Schochet, 2008). It is 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). Holm's approach is less conservative and more powerful than the classical Bonferroni procedure because of the sequentially adjusted *p* values in this method compared to Bonferroni's adjustment fixed across all tests. 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, & Speed, 2007).

The *p* values for the *n* hypotheses being tested are ordered from the smallest to largest ($p_1 \le p_2 \le ... \le p_n$). The smallest *p* value p_1 is tested at α/n (i.e., the Bonferroni adjustment) and if $p_1 > \alpha/n$ then that and all *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 $\alpha/(n-1)$ and if $p_2 > \alpha/(n-1)$ then that and 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 p_i is the *i*th smallest *p* value [or equivalently, until $(n-i+1)p_i > \alpha$]. In the example, only main effects A and B are statistically significant using Holm's procedure because the third smallest unadjusted *p* value (.011 for interaction BC) is larger than the adjusted α of .01 [i.e., .05/(7−3+1), or as shown in Table 2, adjusted *p* = .055, which is larger than α = .05].

Hochberg. Hochberg (1988) proposed another sequentially rejective Bonferroni procedure that is called a "step-up" procedure. This method has shown to be slightly more powerful than the Holm approach based on a Monte Carlo simulation in factorial designs (Kromrey & Dickinson, 1995). However, it lacks stability under certain conditions, for example, when the test statistics are dependent or correlated (Holland & Copenhaver, 1987; Schochet, 2008).

The *p* values of *n* tests are ordered from the largest to the smallest in Hochberg’s method ($p_n \geq p_{n-1} \geq ... \geq p_2 \geq p_1$). The largest *p* value (p_n) is tested at α/1 and if $p_n < \alpha$ then that and all the subsequent tests are rejected; otherwise, the next largest *p* value p_{n-1} is tested at α/2 and if $p_{n-1} \leq \alpha/2$ then that and all the subsequent hypotheses are rejected; otherwise, the third largest *p* value is examined. Generally, if $p_i < \alpha/(n-i+1)$, where *i* is the *i*th smallest *p* value, that and all the subsequent hypotheses are rejected (or equivalently, $(n-i+1)p_i < \alpha$). In fact, the critical values in this procedure are the same as that in Holm’s method---the difference is the order with which the hypotheses are tested. In the example from Table 2, main effects *A*, *B*, and *C* as well as the interaction term *BC* are significant because the fourth smallest unadjusted *p* value (.012 for main effect C) is smaller than the adjusted α = .0125 (i.e., .05/(7−4+1)---or as shown in Table 2, adjusted p = .048, which is less than α = .05.

Benjamini-Hochberg. The Benjamini-Hochberg (1995) method (B-H) focuses on controlling FDR, which also controls FWER in a weak sense. Schochet (2008) reported that this "step-up" procedure has become increasingly popular among researchers. It has been shown to be more powerful than some FWER methods, such as the Bonferroni and the Holm procedures, especially when testing large numbers of hypothesis tests in a family (Keselman et al., 2002; Schochet). Williams et al. (1999) showed that the B-H method is more powerful than the Bonferroni procedure and the Hochberg method. The handbook of the National Assessment of Educational Progress (2009) indicates that the Benjamini-Hochberg FDR procedure is more suitable than other procedures.

The steps of B-H method are (a) to order the *p* values of the *n* tests from the smallest to the largest ($p_1 \leq p_2 \leq ... \leq p_\mathrm{n}$), (b) to find the largest *i*th *p* value which satisfies $p_i \leq (i/n)\alpha$ (or equivalently $np_i/i \leq \alpha$), and (c) to reject all statistical null hypotheses from the first to the *i*th hypothesis (Benjamini & Hochberg, 1995; Cai, 2006; Keselman et al., 2002; Schochet, 2008). Therefore, in this example, the first four hypotheses are rejected: three main effects (*A*, *B*, and *C*) and one two-way interaction *BC* because the *i*th largest unadjusted *p* value that fulfills the $p_i \leq (i/n)\alpha$ condition is the .012 for main effect C (where adjusted α is .029---or equivalently the adjusted *p* is .021, as shown in Table 2).

It should be noted that a number of modified FDR procedures have been developed since 1995. For example, Benjamini and Hochberg (2000) developed an adaptive B-H method for when m_0 is unknown. There are several other procedures which focus on the FDR under different conditions (Benjamini & Liu, 1999; Benjamini & Yekutieli, 2001; Kwong, Holland, & Cheung, 2002; Storey, 2002; Troendle, 2000). However, the original Benjamini and Hochberg (1995) approach is the most commonly applied.

Factorial ANOVA

Even though adjustments for multiple tests are not traditionally made for factorial ANOVA, the familywise Type I error rate inflates in factorial ANOVA because several *F* tests are usually conducted in the same factorial design simultaneously. The Bonferroni method was recommended in a factorial ANOVA design to control the Type I error rate by some researchers (Fletcher, Daw, & Young, 1989; Rosenthal & Rubin, 1984; Smith, Levine, Lachlan, & Fediuk; 2002; Stevens, 2002). Kromrey and Dickinson (1995) found that the Hochberg method performs slightly better than the Holm procedure in terms of power. Indeed, Kromrey and Dickinson determined that Bonferroni, Holm, and Hochberg all controlled the familywise Type I error rates better than the omnibus *F* test (used as protection for other tests) under partial null conditions. Smith et al. recommended the Bonferroni procedure because of its stability in controlling the Type I error rate under both complete null and partial null situations in three-way and four-way balanced factorial designs.

Purpose of the Study

The purpose of this study was to evaluate four MHT procedures (the Bonferroni procedure, the Holm procedure, the Hochberg procedure, and the Benjamini-Hochberg procedure) in terms of Type I error rates in the balanced two-way, three-way, and four-way factorial ANOVA designs. No studies of Benjamini-Hochberg within factorial ANOVA had yet been conducted. The second purpose was to investigate the statistical power rate of these four MHT procedures in the balanced two-way and three-way factorial ANOVA. For comparison, the Type I error rates and statistical power of the four MHT procedures were also compared to those rates obtained in the unadjusted alpha per test procedure, in which each *p* value from each test is compared to .05.

Very few, if any, researchers have investigated both FWER and FDR in the same study (none could be found). Because there is a functional relationship between Type I error and statistical power, it is important to understand the Type I error rates of FDR approaches (e.g., B-H). Before comparing the power rates of B-H with the Holm and Hochberg methods, we need to understand whether the higher power that has been found for B-H is related to an inflated FWER. It should also be noted that results from these balanced factorial ANOVA designs will generalize to other such independent analyses.

Methodology

A Monte Carlo program was created using the computer statistical program *R*, which 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 are performed using built-in *R* functions with the requirement of some additional *R* packages.

Two-way (2x2), three-way (2x2x2), and four-way (2x2x2x2) factorial ANOVA designs were studied. One dependent variable and two, three, or four factors were generated in the two-way, three-way, and four-way designs, respectively. Sample size was 32 per cell in the two-way and 16 per cell in the three-way and four-way balanced designs. All statistical analyses were done with at least 10,000 replications as a Monte Carlo simulation in the R programming language.

Normally distributed dependent variables were generated in each cell using the function *rnorm* in R with cell means set for each pattern so that the necessary number of null hypotheses were true. In a multifactor ANOVA design, the number of true null hypotheses (m_0) will vary in different scenarios. Several patterns of true null and false null effects were studied: 4 patterns in the two-way, 8 patterns in the three-way, and 16 patterns in the four-way design. All non-redundant patterns were used. That is, because the tests are considered independent in a balanced factorial ANOVA, only one pattern with two true null hypotheses is required, no matter which effects are the true null hypotheses. Therefore, for example, only four patterns are required in the two-way design to represent 0, 1, 2, and 3 true null hypotheses for the main effects and/or interaction effect. Where any null hypotheses were set as false for the population, medium effect sizes (.50) were used between the contrasted groups.

The cell means used for the patterns are shown in Table 3 and Table 4. All standard deviations were 1.0. Different cell mean patterns produce different numbers of non-null (or true-null) effects. For example, there are three hypothesis tests in the two-way balanced design: two main effects (A and B) and one interaction effect (AB). For example, the cell means in pattern one are all equal to zero and therefore represent the completely null case where none of the effects are non-null. The cell means in pattern two are 0.0, 0.5, 0.0, and 0.5, resulting in true null hypotheses for one main effect and the interaction; the second main effect is therefore a false null hypothesis for the population (i.e., a real difference).

The factorial ANOVA analyses were conducted in R using built-in functions. Significance (*p*) values were obtained from each analysis and then adjusted as required for each adjustment procedure (Bonferroni, Holm, Hochberg, and Benjamini-Hochberg) were calculated in R using the function *p.adjust*. Familywise alpha was set at .05 for all analyses.

Three Type I error rates were studied: Samplewise familywise error rate (SFWER), Testwise FWER (TFWER), and false discovery rate (FDR). SFWER is measured as the proportion of samples in which any Type I error occurred. Although some researchers might call SFWER as defined here "experimentwise error rate," not all scholars agree that experimentwise error rate is the same as familywise error rate in multifactor designs (e.g., Maxwell & Delaney, 2000; Ryan, 1959). The argument

Table 3. Patterns of Means for 2-way Balanced Design

Number of True Null	Pattern Numbers	Cell (1,1)	Cell (1,2)	Cell (2,1)	Cell (2,2)	True Effect
3	1	0.00	0.00	0.00	0.00	A, B, AB
2	2	0.00	0.50	0.00	0.50	A, AB
1	3	0.50	1.00	0.00	0.50	AB
0	4	0.00	0.00	1.00	0.00	None

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

Table 4. Patterns of means for the 3-way Balanced Design

# of True Null	Pattern	Cell 1,1,1	Cell 1,2,1	Cell 2,1,1	Cell 2,2,1	Cell 1,1,2	Cell 1,2,2	Cell 2,1,2	Cell 2,2,2	True Effect
7	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	all
6	2	0.0	0.5	0.0	0.5	0.5	0.0	0.5	0.0	A,B,C,AB,AC,ABC
5	3	1.0	0.5	0.5	0.0	0.5	0.0	1.0	0.5	A,C,AB,BC,ABC
4	4	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	A,AB,AC,ABC
3	5	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	B,AC,ABC
2	6	0.5	0.5	0.0	−1.0	1.0	0.0	0.5	0.5	A,ABC
1	7	0.5	0.5	1.0	1.0	1.0	−1.0	0.5	0.5	AC
0	8	1.0	0.0	0.0	1.0	−1.0	0.0	0.0	1.0	None

Note. *n* per cell is 16. Pattern 1 for Type I error rate and pattern 2-7 for statistical power.

depends on the definition of family, that is, whether there is a single or several families in one factorial experiment. Here, "samplewise" defines family based on the total number of null hypotheses tested within the sample (e.g., Games, 1971; Ryan, 1959; Stevens, 1999). Testwise FWER measured the rate of Type I errors committed at the test level. TFWER is defined as the proportion of all null hypothesis tests that resulted in a Type I error, resulting in an average per test Type I error rate. Within any one sample, there may be multiple true null hypotheses and therefore multiple possible errors; TFWER uses the count of all such errors in its calculation. False Discovery Rate (FDR) measured the proportion of total rejections that were Type I errors. Whereas TFWER is the proportion of false rejections out of the total number of true null hypotheses, FDR is the proportion of the false rejections out of the total number of rejections. The total number of rejections includes both correct and incorrect rejections of the null hypotheses. Using a somewhat parallel approach, three criteria for statistical power rates were employed: samplewise power, testwise power, and true discovery rate. Samplewise Power (SPOWER) measured the number of samples that resulted in any correctly rejected null hypotheses (no power rate was calculated for samples in which all false null hypotheses were correctly rejected). More specifically, SPOWER is defined as the probability of detecting at least one significant effect in samples that have more than one false null hypothesis. This definition follows from Kirk's (1995) statement that "power refers to the probability of rejecting a false null hypothesis" (p. 58), rather than a definition based on Type II error. Testwise Power (TPOWER) provided a measure of correct rejections at the individual hypothesis test level. TPower is defined as the probability of detecting significant results from among all false null hypotheses, which can be considered the average power based for all false null tests. True Discovery Rate (TDR) provided the proportion of all rejections that were correctly rejected. These definitions mirror those discussed by others in the context of MCPs (e.g., Kirk; Kromrey & Dickinson, 1995; Toothaker, 1993): any-pairs power, all-pairs power, and per-pair power.

Results

Summary of Type I Error Rates

All MHT procedures were found to have advantages over the unadjusted alpha per test procedure in terms of controlling the Type I error inflation. Both the Holm and Hochberg procedures were able to control the Type I error rates at .05, with Hochberg performing slightly better in this regard. The Bonferroni procedure was able to control Type I error, but was much more conservative than the other approaches. None of these results are surprising when put into context of the literature. These results held up across the two-way, three-way, and four-way ANOVA designs and can be seen in Tables 5-7 and Figures 2-7.

However, it is interesting to note that while the Holm and Hochberg procedures are able to control FWER within a range near .05, and Bonferroni was able to keep FWER below .05, there was no such "control" for FDR. That is, no method controlled FDR within range of or below a certain given value (see Figure 3 and Figure 6). Further, all methods designed to control FWER also controlled FDR. Interestingly, the B-H approach allowed more false discoveries than Holm, Hochberg, or Bonferroni.

The most important result---which was not found elsewhere in the literature---is that the SFWER from the B-H procedure became inflated under certain conditions, sometimes severely so (see Tables 5, 6, and 7, and Figures 2, 4, and 7). Specifically, in the middle numbers of true null hypotheses (i.e., 2 true in the two-way design, 3-5 true in the three-way design, and 5-11 true in the four-way design), SFWER increased above the nominal familywise alpha level. For example, while actual SFWER was only .058 in the two-way analysis, it increased to .091 in the 4 true condition of the three-way design and increased to an incredible .192 in the four-way design. This result was not found in the extant literature. These results for SFWER for B-H calls into serious question any claims about higher power for that procedure. It was true, however, that the B-H controlled SFWER in the weak sense (when the completely null was true or almost true), and also when very few null hypotheses were true (e.g., only 1-2 null hypotheses were true in the population).

Summary of Statistical Power Rates

Similar to the results for the Type I error rates, the results presented here for statistical power generally supported what has been reported in the literature. The B-H procedure showed the highest power of the MHT procedures in most situations. Hochberg showed a very slight power advantage over Holm for both SPOWER and TPOWER. Bonferroni had the lowest power of the MHT procedures across all conditions. These results can be seen in Tables 8-9 and Figures 8-10.

What has not been reported previously is the TDR results presented in Tables 8-9 and Figure 10. Because the B-H procedure results in the highest level of FDR, it also results in the lowest level of true discoveries (TDR) among the MHT procedures.

Most importantly, perhaps, is the supplemental analysis presented in Figure 11. These results were created using a slightly relaxed nominal alpha for Bonferroni, Holm, and Hochberg ($\alpha = .07$) while using $\alpha = .05$ for the Benjamini-Hochberg procedure. This supplemental Monte Carlo was performed to investigate the higher actual SFWER found for the B-H procedure above. Clearly, when nominal alpha is relaxed just a little (from .05 to .07), the SPOWER for both Holm and Hochberg procedures exceeds the power for the B-H approach. Indeed, even the SPOWER for the Bonferroni technique exceeds the B-H power when ($\alpha = .07$) is used as familywise nominal alpha for the Bonferroni approach.

Table 5. Type I Error Rates in 2-Way Balanced Design with 20,000 Replications

Number of True Null H_0	Error Rates	Unadjusted	Bonferroni	Holm	Hochberg	B-H
	SFWER	0.1393	0.0476	0.0476	0.0478	0.0483
3	TFWER	0.0486	0.0162	0.0164	0.0166	0.0172
	FDR	1.0000	1.0000	1.0000	1.0000	1.0000
	SFWER	0.0974	0.0329	0.0440	0.0456	0.0578
2	TFWER	0.0499	0.0166	0.0228	0.0239	0.0301
	FDR	0.1106	0.0479	0.0645	0.0674	0.0829
	SFWER	0.0509	0.0166	0.0371	0.0416	0.0438
1	TFWER	0.0509	0.0166	0.0371	0.0416	0.0438
	FDR	0.0307	0.0124	0.0260	0.0288	0.0291

Table 6. Type I Error Rates in 3-Way Balanced Design with 20,000 Replications

Number of True Null H_0	Error Rates	Unadjusted	Bonferroni	Holm	Hochberg	B-H
7	SFWER	0.2913	0.0479	0.0479	0.0479	0.0491
	TFWER	0.0492	0.0071	0.0071	0.0071	0.0078
	FDR	1.0000	1.0000	1.0000	1.0000	1.0000
6	SFWER	0.2627	0.0400	0.0443	0.0446	0.0682
	TFWER	0.0504	0.0068	0.0077	0.0077	0.0125
	FDR	0.2731	0.0702	0.0782	0.0787	0.1203
5	SFWER	0.2217	0.0360	0.0427	0.0431	0.0809
	TFWER	0.0493	0.0073	0.0088	0.0089	0.0175
	FDR	0.1334	0.0326	0.0384	0.0387	0.0674
4	SFWER	0.1883	0.0297	0.0404	0.0411	0.0909
	TFWER	0.0509	0.0075	0.0103	0.0105	0.0244
	FDR	0.0780	0.0182	0.0237	0.0240	0.0473
3	SFWER	0.1396	0.0201	0.0334	0.0346	0.0803
	TFWER	0.0492	0.0068	0.0115	0.0120	0.0283
	FDR	0.0442	0.0093	0.0147	0.0152	0.0298
2	SFWER	0.0934	0.0140	0.0277	0.0298	0.0645
	TFWER	0.0482	0.0070	0.0143	0.0158	0.0335
	FDR	0.0234	0.0051	0.0094	0.0103	0.0182
1	SFWER	0.0511	0.0065	0.0208	0.0270	0.0407
	TFWER	0.0511	0.0065	0.0208	0.0270	0.0407
	FDR	0.0105	0.0020	0.0055	0.0070	0.0090

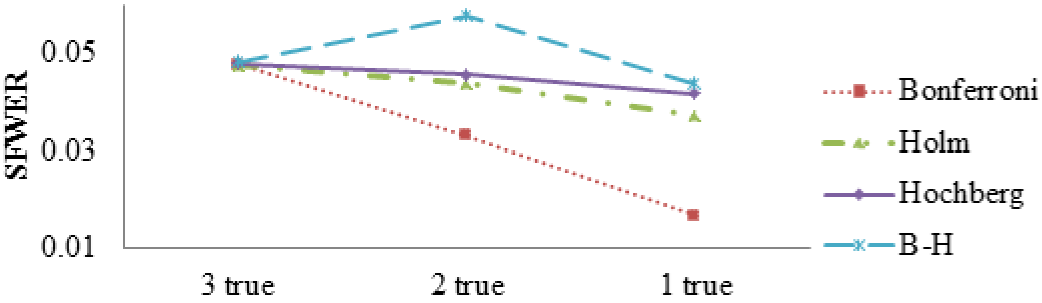

Figure 2. Samplewise familywise error rates in 2-way balanced design with 20,000 replications.

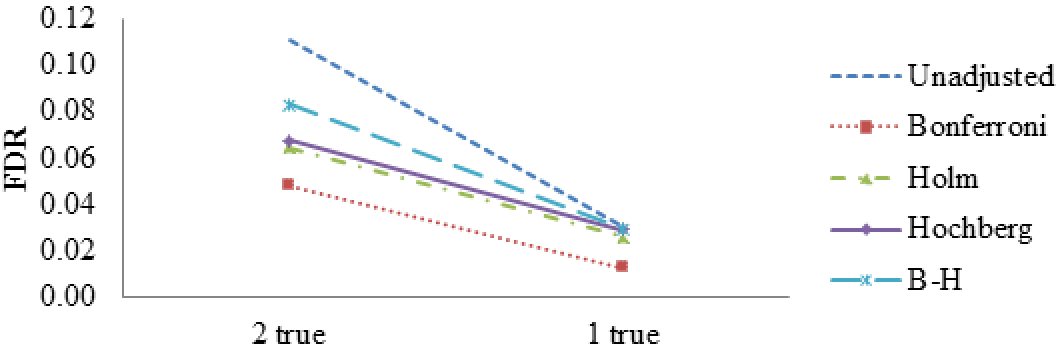

Figure 3. Samplewise false discovery rates in 2-way balanced design with 20,000 replications.

Table 7. Samplewise Type I Error Rates in Four-Way Balanced Design with 20,000 Replications

Number of True Null H_0	Error Rates	Unadjusted	Bonferroni	Holm	Hochberg	B-H
15	SFWER	.5329	.0472	.0472	.0472	.0482
14	SFWER	.5084	.0468	.0496	.0496	.0902
13	SFWER	.4847	.0439	.0508	.0508	.1216
12	SFWER	.4591	.0390	.0488	.0489	.1481
11	SFWER	.4359	.0362	.0496	.0496	.1720
10	SFWER	.3991	.0325	.0483	.0483	.1804
9	SFWER	.3732	.0296	.0484	.0484	.1918
8	SFWER	.3350	.0264	.0473	.0474	.1880
7	SFWER	.3000	.0217	.0478	.0478	.1923
6	SFWER	.2682	.0220	.0520	.0520	.1888
5	SFWER	.2260	.0156	.0420	.0426	.1640
4	SFWER	.1795	.0137	.0376	.0382	.1384
3	SFWER	.1426	.0099	.0336	.0350	.1164
2	SFWER	.0942	.0069	.0200	.0207	.0764
1	SFWER	.0484	.0036	.0097	.0101	.0395

Table 8. Statistical Power Rates in 2-Way Balanced Design with 20,000 Replications

Number of False Null H_0	Power Rates	Unadjusted	Bonferroni	Holm	Hochberg	B-H
1	SPOWER	0.8029	0.6578	0.6600	0.6612	0.6646
	TPOWER	0.8029	0.6578	0.6600	0.6612	0.6646
	TDR	0.8894	0.9521	0.9355	0.9326	0.9171
2	SPOWER	0.9572	0.8784	0.8790	0.8828	0.8878
	TPOWER	0.8029	0.6586	0.6956	0.7011	0.7297
	TDR	0.9693	0.9876	0.9740	0.9712	0.9709
3	SPOWER	0.9919	0.9578	0.9578	0.9629	0.9659
	TPOWER	0.7997	0.6535	0.7474	0.7622	0.7745
	TDR	1.0000	1.0000	1.0000	1.0000	1.0000

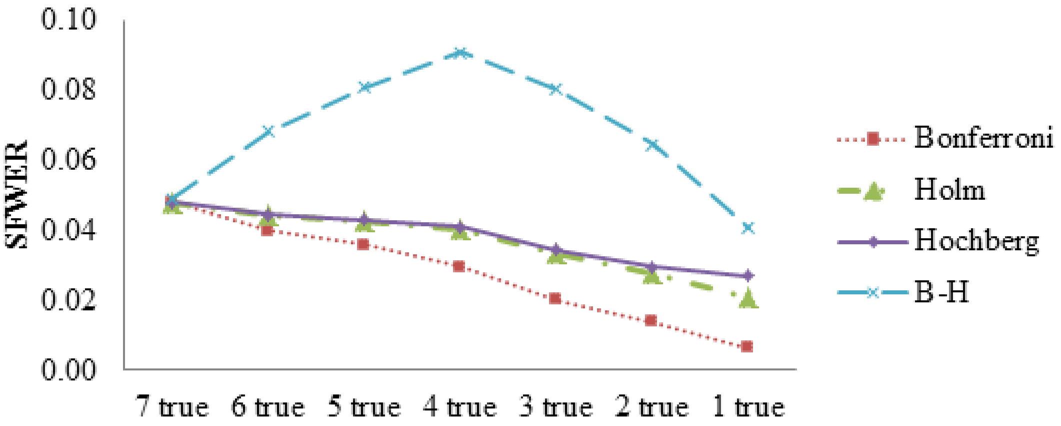

Figure 4. Samplewise familywise error rates in 3-way balanced design with 20,000 replications

Table 9. Statistical Power Rates in 3-Way Balanced Design with 20,000 Replications

Number of False Null H_0	Power Rates	Unadjusted	Bonferroni	Holm	Hochberg	B-H
1	SPOWER	0.8041	0.5407	0.5417	0.5419	0.5493
	TPOWER	0.8041	0.5407	0.5417	0.5419	0.5493
	TDR	0.7269	0.9298	0.9218	0.9213	0.8797
2	SPOWER	0.9607	0.7842	0.7849	0.7856	0.7987
	TPOWER	0.8005	0.5411	0.5523	0.5529	0.6061
	TDR	0.8666	0.9674	0.9616	0.9613	0.9326
3	SPOWER	0.9922	0.8949	0.8952	0.8961	0.9105
	TPOWER	0.8022	0.5408	0.5669	0.5684	0.6540
	TDR	0.9220	0.9818	0.9763	0.9760	0.9527
4	SPOWER	0.9979	0.9469	0.9470	0.9475	0.9596
	TPOWER	0.7980	0.5373	0.5786	0.5814	0.6893
	TDR	0.9558	0.9907	0.9853	0.9848	0.9702
5	SPOWER	0.9997	0.9761	0.9761	0.9766	0.9839
	TPOWER	0.8035	0.5411	0.6011	0.6063	0.7247
	TDR	0.9766	0.9949	0.9906	0.9897	0.9818
6	SPOWER	0.9999	0.9859	0.9859	0.9865	0.9924
	TPOWER	0.8025	0.5395	0.6239	0.6361	0.7495
	TDR	0.9895	0.9980	0.9945	0.9930	0.9910
7	SPOWER	1.0000	0.9940	0.9940	0.9947	0.9969
	TPOWER	0.8033	0.5408	0.6605	0.6894	0.7729
	TDR	1.0000	1.0000	1.0000	1.0000	1.0000

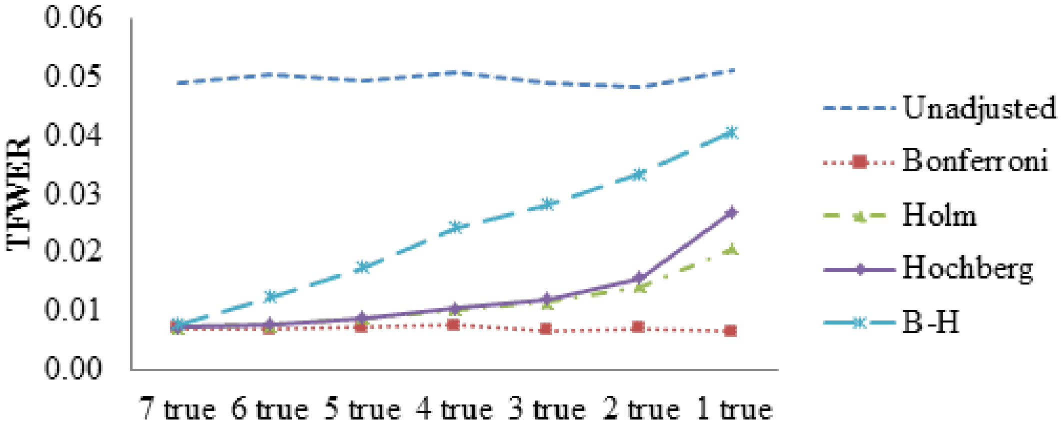

Figure 5. Testwise familywise error rates in 3-way balanced design with 20,000 replications

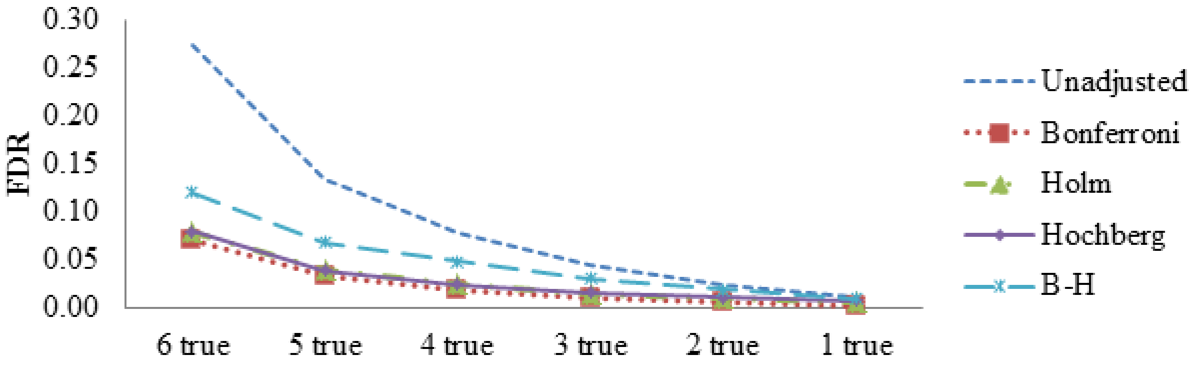

Figure 6. False discovery rates in 3-way balanced design with 20,000 replications.

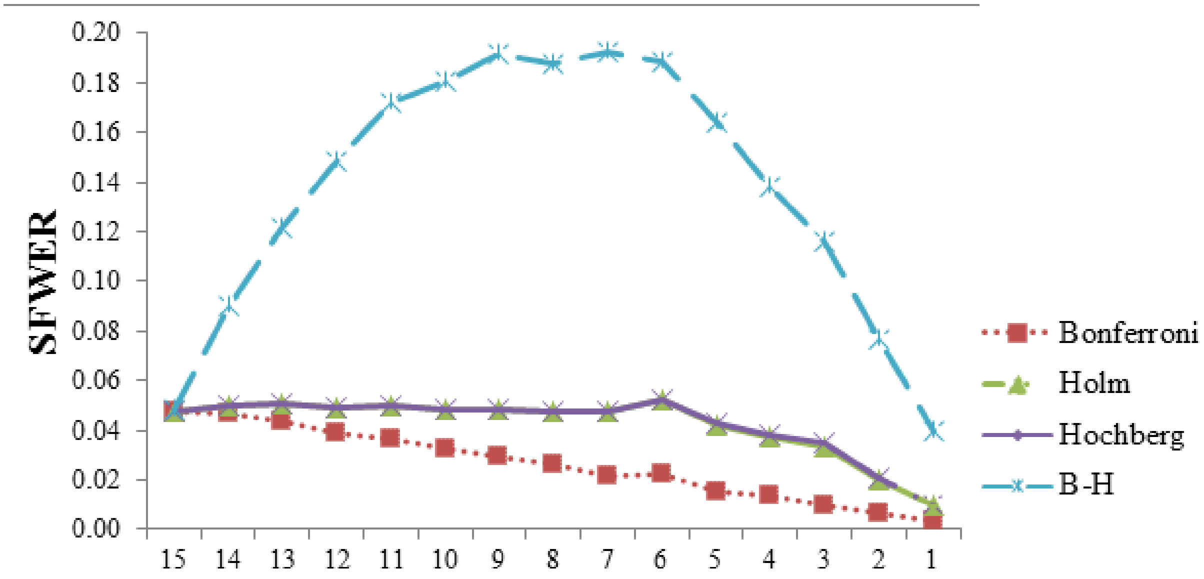

Figure 7. Samplewise familywise error rates in 4-way balanced design with 20,000 replications.

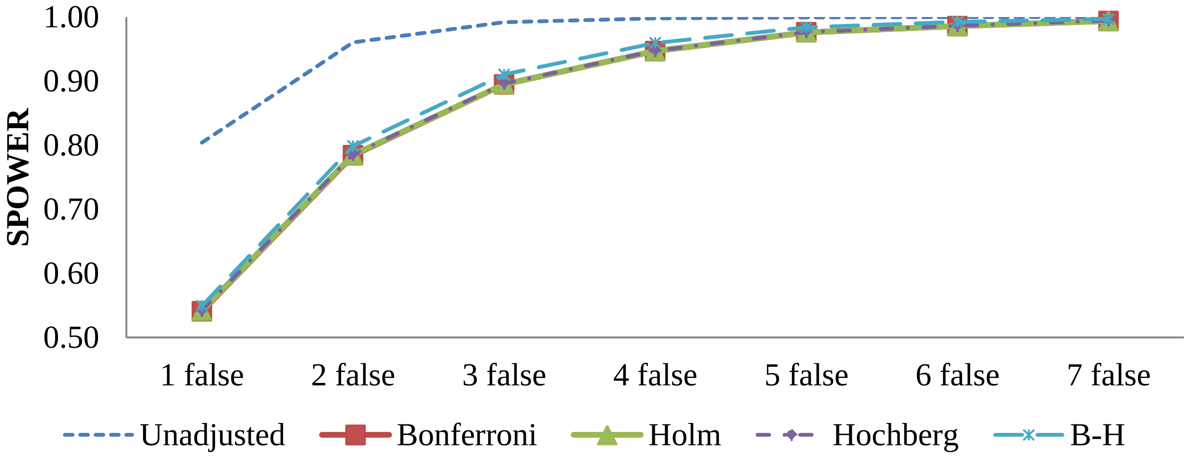

Figure 8. Samplewise statistical power rates in 3-way balanced design with 20,000 replications.

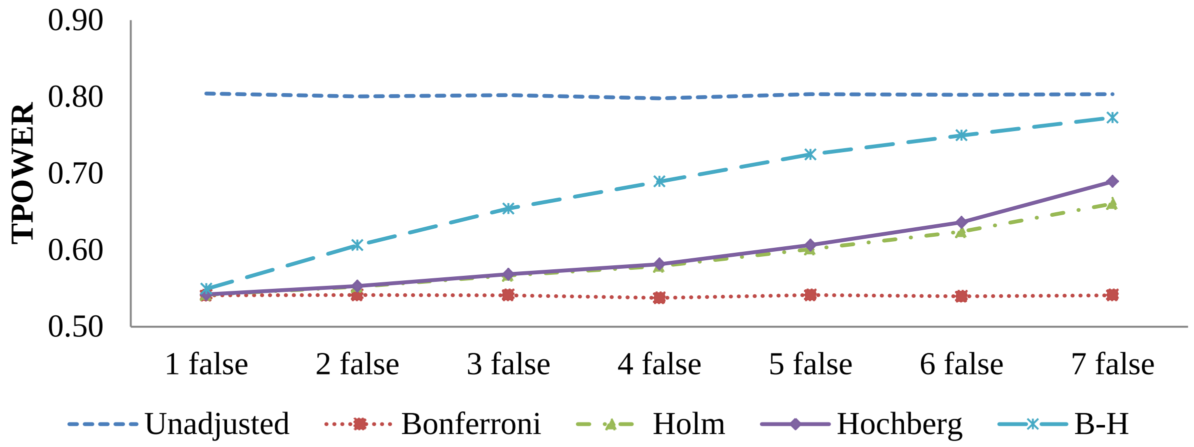

Figure 9. Testwise statistical power rates in 3-way balanced design with 20,000 replications.

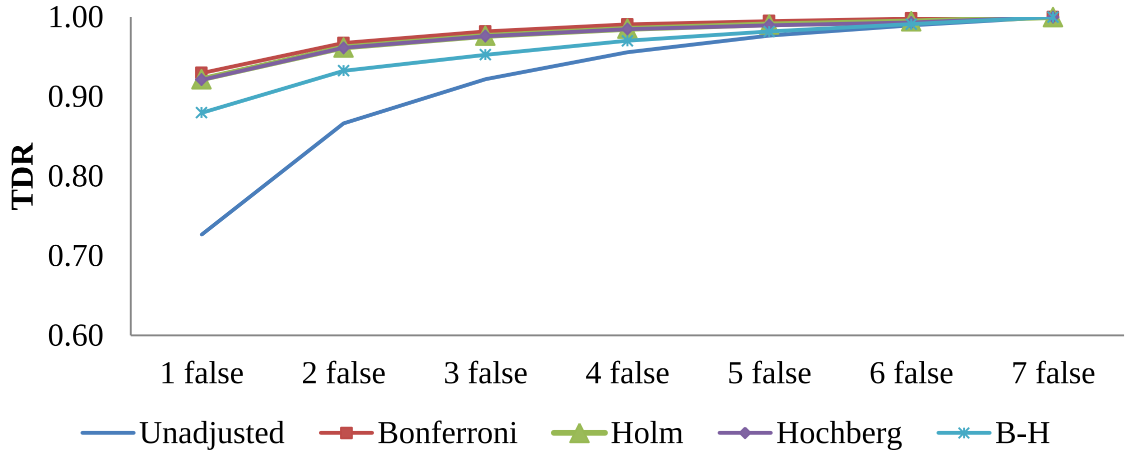

Figure 10. True discovery rates in 3-way balanced design with 20,000 replications.

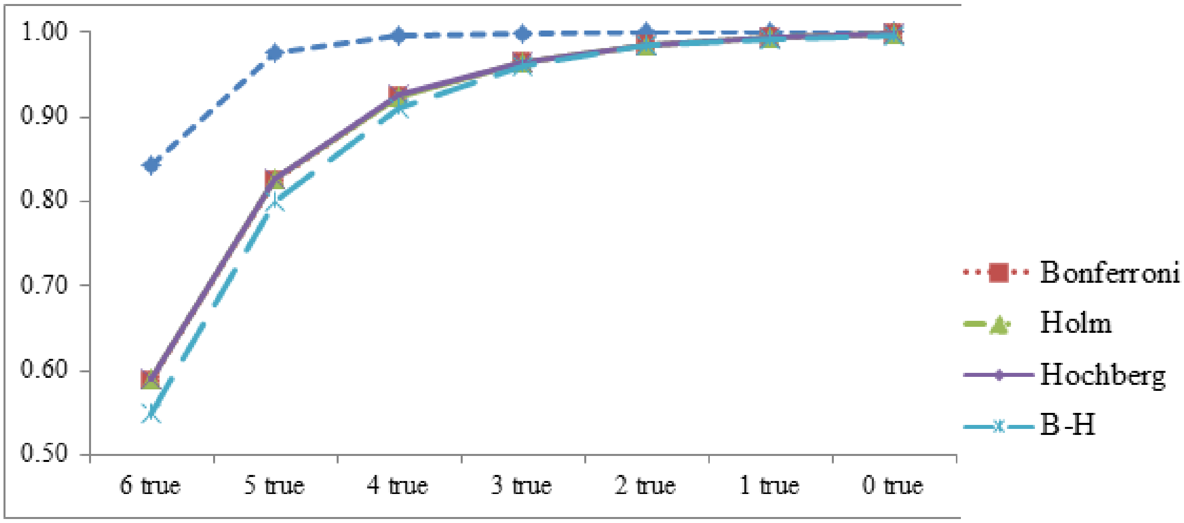

Figure 11. Samplewise statistical power rates in 3-way balanced design with 20,000 replications, with nominal alpha .05 for BH and .07 for other methods.

Conclusions

Being an important statistical design, factorial ANOVA is used broadly in educational research. However, more importantly, the uncorrelated effects in balanced factorial designs allow this study to provide generalizability to other statistical situations that meet the assumption of independence. That is, these results apply to any General Linear Model (GLM) statistical analysis with multiple uncorrelated hypothesis tests. Indeed, some MHT methods (e.g., Hochberg and Benjamini-Hochberg) require independence of tests. The Hochberg procedure worked best in this study overall, slightly better than Holm. However, due to the independence assumption required by Hochberg, the Holm procedure may be the best choice if independence of tests is not certain, which may be true in a number of situations.

Previous literature shows many great contributions to MCPs to control Type I error inflation in one-way ANOVA; however, 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 discussion and consideration of the issue are needed. It is necessary for researchers to consider further Type I error rate inflation in factorial ANOVA (Smith et al., 2002; Stevens, 2002) and other general linear model designs such as tests of multiple regression coefficients (Mundfrom, Perrett, Schaffer, Piccone, & Rozeboom, 2006; Neter et al., 1996) or multiple dependent variables in MANOVA post hoc testing (Stevens, 2002). For example, Mundfrom et al. found Type I error inflation for unadjusted *t* tests used for testing individual regression coefficients, as much as 6 times larger than the nominal alpha when 8 predictors were in the model. They noted, however, that the Bonferroni adjustment was too conservative in the regression scenarios they examined.

The Benjamini-Hochberg (1995) procedure has received positive recommendations in recent literature. Indeed, it showed the greatest power here, but at the risk of higher SFWER. While designed to control FDR rather than to control Type I error, it is not fair to credit the Benjamini-Hochberg procedure with higher power (as many have done) if it comes at the risk of inflated Type I errors. For example, we are able to achieve similar (even higher) power rates with the Holm and Hochberg procedures simply by increasing the nominal alpha to .07, still below the Type I error inflation often obtained using Benjamini-Hochberg. There are philosophical reasons that must be considered in regard to whether to adjust alpha for multiple hypothesis tests; however, these decisions must also be made with informed consideration of the statistical impact affiliated with the procedures.

Researchers should have a better understanding about FDR. There are not many studies that have been found to investigate a good value for FDR (Benjamini & Hochberg, 1995; Williams, et al., 1999). The Type I error should be controlled at .05, but for FDR, it is not clear that the smaller is the better, or that it also should be or can be controlled at a nominal value. Similarly, further consideration of the usefulness of TDR may be warranted. Also, this study did not include a measure of samplewise FDR, which might prove useful as scholars continue to unravel the differences between FWER and FDR.

Investigating unbalanced designs will help determine how well these methods work with correlated tests. We believe that less-sophisticated researchers will apply the Hochberg and Benjamini-Hochberg methods---because of the generally favorable recommendations they have received---even with non-independent hypothesis tests.

Because there are different opinions about how to define family in factorial ANOVA (and indeed, in many other types of statistical analyses), future studies may define the effects as separate families. All hypothesis tests were regarded as a single family in this study, but family can also be defined in terms of main effects and interaction effects separately. Also, the sample size per cell was fixed in this study, but future studies might compare the Type I error rate and statistical power rate with different, unbalanced sample sizes per cell. The assumptions of factorial ANOVA were not violated in this study, so future studies might investigate situations where the assumptions are violated.

Practical Implications and Recommendations

The current study showed that 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 may be concerned about the FWER, as Storey

(2003) has noted, "The FWER offers an extremely strict criterion which is not always appropriate" (p. 2014). Stevens (2002) has suggested using a liberal level of alpha, for example, .10 or .15 in MHT, which is equivalent to an FDR approach with its "weak control" of FWER. That is, the liberal alpha level can lead to results just as powerful, if not more powerful, than those obtained using FDR approaches.

Finally, although scholars have different opinions about adjusting alpha, many have agreed that MHT procedures are useful statistical tools (Hewes, 2003; O'Keefe, 2003; Tukey, 1991). The use of MHT procedures depends on whether researchers choose to adjust alpha. It is a philosophical issue and researchers have to make their own decisions, perhaps based on the context of their relevant literature base. When a researcher can clearly define a family of tests within a study; however, it seems very clear that some sort of alpha-adjustment procedure must be used. 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 (i.e., whether to control FWER or FDR). As Tukey suggested, there is no unique answer toward controlling the Type I error inflation, but the results reported here help inform this conversation.

References

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

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics, 25*, 60-83.

Benjamini, Y., & Liu, W. (1999). A step-down multiple hypothesis testing procedures that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference, 82*, 163-170.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Cai, G. Q. (2006). *Further results on Simes' test and Benjamini-Hochberg False Discovery Rate procedure*. Unpublished doctoral dissertation, Temple University. Retrieved from Dissertations and Theses database.

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

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Thousand Oaks, CA: Sage.

Fletcher, H. J., Daw, H., & Young, J. (1989). Controlling multiple *F* tests errors with an overall *F* tests. *Journal of Applied Behavioral Science, 25*, 101-108.

Games, P. A. (1971). Multiple comparisons of means. *American Educational Research Journal, 8*, 531-565.

Ge, Y., Sealfon, S. C., Tseng, C. & Speed, T. P. (2007). A Holm-type procedure controlling the false discovery rate. *Statistics and Probability Letters, 77*, 1756-1762.

Halderson, J. S., & Glasnapp, D. R. (1974, April). Error rates of multiple *F* tests in factorial ANOVA designs. Paper presented at the annual of meeting of the American Educational Research Association, Chicago, IL.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*(3), 269-306.

Hewes, D. E. (2003). Methods as tools: A response to O'Keefe. *Human Communication Research, 29*, 448-454.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803.

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine, 9*, 811-818.

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

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43*, 417-423.

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

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keren, G., & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Keselman, J. C., Cribbie, R., & Holland, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology, 55*, 27-39.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Kromrey, J. D., & Dickinson, W. B. (1995). The use of an overall *F* test to control Type I error rates in factorial analysis of variance: Limitations and better strategies. *Journal of Applied Behavioral Science, 31*(1), 51-64.

Kwong, K. S., Holland, B., & Cheung, S. H. (2002). A modified Benjamini-Hochberg multiple comparisons procedure for controlling the false discovery rate. *Journal of Statistical Planning and Inference, 104*, 351-362.

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

Maxwell., S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

Miller, R. G. (1981). *Simultaneous statistical inference*. New York: Wiley.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Newbury Park, CA: Sage.

Mundfrom, D. J., Perrett, J. J., Schaffer, J., Piccone, A., & Rozeboom, M. (2006). Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints, 32*, 1-6.

National Center for Education Statistics (2009). Statistical standards. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/infer_multiplecompare.asp

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wassermen, W. (1996). *Applied linear statistical models* (4th ed.). Boston: McGraw Hill.

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

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*, *76*, 1028-1034.

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

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin, 57,* 318-328.

Schochet, P. Z. (2008). Guidelines for multiple testing in impact evaluations of educational interventions. *Mathematica Policy Research, Inc.*, 1-34.

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

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and Type II error rates in multiway ANOVA. *Human Communication Research, 28*, 515-530.

Stevens, J. (1999). *Intermediate statistics: A modern approach* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B, 64*, 479-498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics, 31*, 2013-2035.

Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage.

Troendle, J. F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference, 84*, 139-158.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100-116.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

What Works Clearinghouse (2008, December). Procedures and standards handbook. Retrieved from http://ies.ed.gov/ncee/wwc/

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42-69.

Send correspondence to:

Gordon P. Brooks
Ohio University
Email: brooksg@ohio.edu