Hidden Multiplicity in Multiway ANOVA: Prevalence, Consequences, and Remedies

Angélique O. J. Cramer

Don van Ravenzwaaij

Dora Matzke

Helen Steingroever

Ruud Wetzels

Raoul P. P. P. Grasman

Lourens J. Waldorp

Eric-Jan Wagenmakers

Send correspondence to:

Angélique O. J. Cramer
University of Amsterdam
Department of Psychology
Psychological Methods
Weesperplein 4
1018 XA Amsterdam
angecramer@gmail.com
www.aojcramer.com

Abstract

Many empirical researchers do not realize that the common multiway analysis of variance (ANOVA) harbors a multiple comparison problem. In the case of two factors, three separate null hypotheses are subject to test (i.e., two main effects and one interaction). Consequently, the probability of a Type I error is 14% rather than 5%. We explain the multiple comparison problem and demonstrate that researchers almost never correct for it. We describe one of several correction procedures (i.e., sequential Bonferroni), and show that its application alters at least one of the substantive conclusions in 45 out of 60 articles considered. An alternative method to combat the multiplicity in multiway ANOVA is preregistration of hypotheses.

Keywords: sequential Bonferroni, multiway ANOVA, multiple comparison problem, Type I error, Bonferroni-Holm, preregistration

The factorial or multiway analysis of variance (ANOVA) is one of the most popular statistical procedures in psychology. Whenever an experiment features two or more factors, researchers usually apply a multiway ANOVA to gauge the evidence for the presence of each of the separate factors as well as their interactions. For instance, consider a response time experiment with a 2x3 balanced design (i.e., a design with equal number of participants in the conditions of both factors); factor A is speed-stress (high or low) and factor B is the age of the participants (14-20 years, 50-60 years, and 75-85 years). The standard multiway ANOVA tests whether factor A is significant (at the .05 level), whether factor B is significant (at the .05 level) and whether the interaction term A*B is significant (at the .05 level). In the same vein, the standard multiway ANOVA is also frequently used in non-experimental settings (e.g., to assess the potential influence of gender and age on major depression).

Despite its popularity, few researchers realize that the multiway ANOVA brings with it a problem of multiple comparisons, particularly when detailed hypotheses have not been specified a priori (to be discussed in more detail later). For the 2x3 scenario discussed above without a-priori hypotheses (i.e., when the researcher's attitude can best be described by "let us see what we can find"; de Groot, 1969), three independent tests are carried out, and this means that, in case the data originate from the null hypothesis, there is a probability of 1 − (1 − .05)^3 = .14 of at least one significant result. This is called a Type I error or familywise error rate (FWE). The problem of Type I error is not trivial: add a third, balanced factor to the 2x3 scenario (e.g., a 2x3x3 design), and the probability of finding at least one significant result when H_0 is true increases to 1 − (1 − .05)^7 = .30. The situation becomes even worse when designs have *unequal numbers of participants* per condition: in such unbalanced designs, the three tests in our hypothetical 2x3 experiment are no longer

independent and this further increases the probability of a Type I error (Rao & Toutenburg, 1999). Thus, in the absence of strong a priori expectations about the tests that are relevant, this alpha-inflation is dramatic and should be cause for great concern.

Here we underscore the problem of multiple comparisons inherent in multiway ANOVA. We conduct a literature review and demonstrate that the problem is widely ignored: recent articles published in leading psychology journals contain virtually no procedures to correct for the multiple comparison problem. Next we outline one possible remedy, the sequential Bonferroni procedure (Hartley, 1955; Hochberg, 1988; Holm, 1979; McHugh, 1958; Shaffer, 1986; Wright, 1992). Finally, we demonstrate that the sequential Bonferroni procedure alters at least one of the substantive conclusions in 45 of 60 randomly chosen articles. In order to prevent the loss of power that is inherent to all correction procedures we recommend preregistration of the hypotheses of interest.

Type I Errors and the Oneway ANOVA

A Type I error occurs when a null hypothesis (H_0) is falsely rejected in favor of an alternative hypothesis (H_1). With a single test, such as the oneway ANOVA, the probability of a Type I error can be controlled by setting the significance level α. For example, when α = .05 the probability of a Type I error is 5%. Since the oneway ANOVA comprises only one test, there is no multiple comparison problem. It is well-known, however, that this problem arises in the oneway ANOVA whenever the independent variable has more than two levels and post-hoc tests are employed to determine which condition means differ significantly from one another. For example, consider a researcher who uses a oneway ANOVA and obtains a significant effect for Ethnicity on the total score of a depression questionnaire. Assume that Ethnicity has

three levels (e.g., Caucasian, African-American, and Asian); then this researcher will usually perform multiple post-hoc tests to determine which ethnic groups differ significantly from one another – here the three post-hoc tests are Caucasian vs. African-American, Caucasian vs. Asian, and African-American vs. Asian. Note that when the three test statistics are independent (as is the case when the design is balanced), the overall Type I error equals 1 − (1 − .05)^3 = .14. That is, the probability that at least one post-hoc test leads to a false rejection of H_0 has increased almost threefold. Fortunately, for the oneway ANOVA the multiple comparison problem has been thoroughly studied. Software programs such as SPSS explicitly address the multiple comparison problems by offering a host of correction methods including Tukey's HSD test, Hochberg's GT2, and the Scheffé method (Hochberg, 1974; Scheffé, 1953; Tukey, unpublished).

The Explorative Multiway ANOVA: A Family of Hypotheses

Now consider a design that is only slightly more complicated. Suppose a researcher wants to test whether both Gender (G; two levels) and Ethnicity (E; three levels) influence the total score on a depression questionnaire. Furthermore, suppose that this researcher has no firm a priori hypothesis about how G and E influence the depression total score; that is, the researcher is predominantly interested in finding out whether *any* kind of relationship exists between G, E and depression: a classic example of the *guess* phase of the empirical cycle in which hypotheses are formed rather than tested (de Groot, 1969).

In this case, the multiway ANOVA with two factors, G and E, is an *explorative* one: without strictly formulated a priori hypotheses, the researcher, for example in SPSS, obtains the results for all three hypotheses involved (i.e., main effect of G, main effect of E and a GxE interaction) by means of a single mouse click. As such, in

an explorative setting, all hypotheses implied by the design are considered and tested jointly, rendering this collection of hypotheses a *family*; in line with the idea that "…the term 'family' refers to the collection of hypotheses […] that is being considered for joint testing" (Lehmann & Romano, 2005). As a result, we argue that a multiple comparison problem lurks in these explorative uses of a multiway ANOVA.

To see this, consider the results of a fictitious explorative multiway ANOVA as reported in **Table 1**. When interpreting the ANOVA table, most researchers would conclude that both main effects as well as the interaction are significant as all *p*-values are smaller than α = .05. This conclusion is intuitive and directly in line with the numbers reported in **Table 1**. Nevertheless, this conclusion is incorrect; because the researcher does not have firm a priori hypotheses and therefore tests all three hypotheses simultaneously; and is therefore engaged in an explorative "fishing expedition". The tests in the multiway ANOVA for balanced designs are independent (Toutenburg, 2002) and thus the multiple comparison problem, when unaddressed, results in a 14% Type I error probability. Note that multiway ANOVAs in the psychological literature often consist of three or four factors and this compounds the problem. In the case of three factors and without a priori hypotheses, the total number of tests is seven (i.e., three main effects, three first-order interactions, and one second-order interaction, 2^3 − 1) and the resulting probability of a Type I error 30% (i.e., 1 − (1 − .05)^7); with four factors, the probability of incorrectly rejecting one or more null hypotheses is 54%. It is therefore incorrect to compare each of the *p*-values from a multiway ANOVA table to α = .05.

This is notably different from the situation where the researcher uses a multiway ANOVA for *confirmatory* purposes; that is, where the researcher tests one or more a priori postulated hypotheses (i.e., hypothesis testing in the *predict* phase of

the empirical cycle; de Groot, 1969). In the case of one predefined hypothesis in a design with two factors, for example, the family is no longer defined as encompassing all hypotheses implied by the design (i.e., three); but as all to-be-tested hypotheses, in this case: one, rendering it unnecessary to adjust the level of α.

Table 1.

Example ANOVA table for the three tests associated with a 2x3 design with Gender (G) and Ethnicity (E) as independent factors.

		df1	*df2*	*F*	*p*-value
Main effect	G	1	30	5	.0329*
	E	2	30	4	.0288*
Interaction	G x E	2	30	4.50	.0195*

*significant at α = .05

The realization that explorative multiway ANOVAs inherently contain a multiple comparison problem may come as a surprise to many empiricists, even to those who use the multiway ANOVA on a regular basis. In standard statistical textbooks, the multiple comparison problem is almost exclusively discussed in the context of one-way ANOVAs. In addition, statistical software packages such as SPSS do not present the possible corrective procedures for the multiway case, and this invites researchers to compare each of the p-values to α = .05.

We are not the first to identify the multiplicity problem (e.g., Didelez, Pigeot & Walter, 2006; Fletcher, Daw & Young, 1989; Kromrey & Dickinson, 1995; Olejnik, Li & Supattathum, 1997; Ryan, 1959; Smith, Levine, Lachlan & Fediuk, 2002). Earlier work on the problem, however, does not feature in mainstream statistical textbooks. Moreover, the majority of this work is written in a technical style that is inaccessible to

scholars without sophisticated statistical knowledge. Consequently, empirical work has largely ignored the multiplicity problem. As we will demonstrate shortly, the ramifications can be profound.

One may argue that the problem sketched above is less serious than it appears. Perhaps the majority of researchers in psychology test a single pre-specified hypothesis, thereby circumventing the multiple comparison problem. Or perhaps, whenever they conduct multiple tests, they use some sort of procedure to adjust the α level for each test. This is not the case. Pertaining to the former, it is unfortunately quite common to perform what Gigerenzer (2004) has termed the "null ritual" in which a researcher specifies H_0 in purely statistical terms (e.g., equality of means) without providing an alternative hypothesis in substantive terms (e.g., women are more depressed than men). Additionally, Kerr (1998) notes that researchers in psychology are quite commonly seduced into presenting a post hoc hypothesis (e.g., Caucasian people are more depressed than African-American people: main effect of ethnicity on depression) as if it were an a priori hypothesis (Hypothesizing After the Results are Known: HARKing). Hence, hindsight bias and confirmation bias make it difficult for researchers to ignore the presence of unexpected “significant” effects (i.e., effects for which the individual test has $p < .05$).

In the next section we address the empirical question of whether researchers correct for multiple comparisons when they use the multiway ANOVA. The short answer is that, almost without exception, researchers interpret the results of the individual tests in isolation, without any correction for multiple comparisons.

Practice: Multiway Corrections in Six Journals

We selected six journals that rank among the most widely read and cited journals in experimental, social, and clinical psychology. For these journals we specifically investigated all 2010 publications:

1. *Journal of Experimental Psychology General*: volume 139, issues 1-4 (40 papers).
2. *Psychological Science*: volume 21, issues 1-12 (285 papers).
3. *Journal of Abnormal Psychology*: volume 119, issues 1-4 (88 papers).
4. *Journal of Consulting and Clinical Psychology*: volume 78, issues 1-6 (92 papers).
5. *Journal of Experimental Social Psychology*: volume 46, issues 1-6 (178 papers).
6. *Journal of Personality and Social Psychology*: volumes 98 and 99, issues 1-6 (136 papers).

For each article, we assessed whether a multiway ANOVA was used. If so, we investigated whether the authors had used some sort of correction procedure (e.g., an omnibus test) to remedy the multiple comparison problem. The results are summarized in **Table 2**.

Table 2.

Percentage of articles overall and in the six selected journals that used a multiway ANOVA, and the percentage of these articles that used some sort of correction procedure

	% papers using mANOVA	% papers using mANOVA + correction
Overall	47.62	1.03

JEPG	84.61	0
Psych Sci	43.16	0
J Abn Psych	31.82	0
JCCP	16.30	0
JESP	65.17	2.59
JPSP	54.41	1.35

Overall, all papers from the six journals together; *JEPG*, Journal of Experimental Psychology: General; *Psych Sci*, Psychological Science; *J Abn Psych*, Journal of Abnormal Psychology; *JCCP*, Journal of Consulting and Clinical Psychology; *JESP*, Journal of Experimental Social Psychology; *JPSP*, Journal of Personality and Social Psychology; mANOVA, multiway ANOVA.

Two results stand out. First, almost half of all articles under investigation here used a multiway ANOVA, underscoring the popularity of this testing procedure. Second, only around 1% of these papers used a correction procedure. In all four cases where a correction procedure was used, this was an omnibus *F*-test. In such a test, one pools the sums of squares and degrees of freedom for all main effects and interactions into a single *F* statistic. The individual *F* tests should only be conducted if both this omnibus H_0 is rejected as well as all other combinations of null hypotheses (Fletcher, Daw & Young, 1989; Wright, 1992). So for example, in the case of a 2x2 ANOVA, one should first test the omnibus hypothesis with all three hypotheses included (two main effects and an interaction). If significant, then one proceeds to test the three combinations of two null hypotheses (i.e., main effects A and B, main effect A and interaction, main effect B and interaction). Finally, if significant, only then can one safely continue and test the individual hypotheses. When this closed test procedure is followed, one is safeguarded against capitalization on chance both in the case of unbalanced and balanced designs (Shaffer, 1995).

In sum, our literature review confirms that the multiway ANOVA is a highly popular statistical method in psychological research, but that its use is almost never accompanied by a correction for multiple comparisons. Note that this state of affair is different for fMRI and genetics research where the problem is more evident and it is common practice to correct for multiplicity (e.g., Poldrack et al., 2008).

Possible Remedy: Correction by Sequential Bonferroni

As noted earlier, some statisticians have been aware of the multiple comparison problem in multiway ANOVA. However, our literature review demonstrated that this awareness has not resonated in the arena of empirical research in psychology.

In the few cases where a correction procedure was used, this involved an omnibus *F* test, a test that cannot control the familywise Type I error under partial null conditions (Kromrey & Dickinson, 1995). For example, suppose that in a three-way ANOVA a main effect is present for one factor but not in the remaining two factors; then the overall *F* test is likely to yield a significant *F* value because, indeed, the omnibus null hypothesis is false. However, the omnibus test does not remedy the multiple comparison problem involving the remaining two factors.

A more general correction is known as the sequential Bonferroni procedure (also known as the Bonferroni-Holm correction), which was first introduced by Hartley (1955) and subsequently (independently) re-invented and/or modified by others (Hochberg, 1988; Holm, 1979; McHugh, 1958; Shaffer, 1986; Rom, 1990; Wright, 1992). How does the procedure work? Let us revisit our hypothetical example in which a researcher conducts a two-way ANOVA with G and E as independent factors (uncorrected results are listed in **Table 1**). The results of the sequential Bonferroni correction procedure for this example are presented in **Table 3**. First, one sorts all

significant *p*-values in ascending order, that is, with the smallest *p*-value first. Next, one computes an adjusted α level, α_{adj}. For the smallest *p*-value, α_{adj} equals α divided by the number of tests. Thus, in this example, we conduct three tests so α_{adj} for the smallest *p*-value equals .05/3 = .01667. For the second *p*-value, α_{adj} equals α divided by the number of tests minus 1. So, in our example, the next α_{adj} equals .05/2 = .025. For the final *p*-value, α_{adj} equals α divided by 1 (i.e., the total number of tests minus 2). So, in our example, the final α_{adj} equals .05/1 = .05. Next, one evaluates each *p*-value against these adjusted α levels, sequentially, with the smallest p-value evaluated first. Importantly, if the H_0 associated with this *p*-value is not rejected (i.e., $p > \alpha_{adj}$) then all testing ends and all remaining tests are considered non-significant as well.

In our example, we evaluate $p = .0195$ against $\alpha_{adj} = .01667$: $p > \alpha_{adj}$ and therefore we conclude that the G x E interaction is not significant. We therefore stop testing and conclude that the remaining main effects are not significant as well. Thus, with the sequential Bonferroni correction procedure, we conclude, for this example, that none of the effects are significant; without a correction procedure, we had concluded that all of the effects were significant.

We note that other correction procedures are available, for example those that focus on the *false discovery rate* (FDR: Benjamini & Hochberg, 1995); these other procedures might result in a different conclusion. The FDR, for example, which we will later discuss in more detail, would have resulted in more effects being judged significant because, relative to the sequential Bonferroni correction, the FDR is less conservative.

Table 3.

The sequential Bonferroni procedure for the hypothetical example of Table 1. The procedure entails: (1) sorting *p*-values in ascending order; (2) computing adjusted α level per test (α_{adj}); (3) sequentially evaluating each *p*-value against these adjusted α levels (i.e., reject or not reject H_0); and (4) stopping whenever H_0 is not rejected (and do not reject all remaining untested H_0).

Effect	*p*-value	α_{adj}	H_0
G x E	.0195	.0167	not rejected
E	.0288	.0250	not rejected
G	.0329	.0500	not rejected

*significant at α_{adj}

Thus, the sequential Bonferroni correction procedure allows one to control for the FWE by evaluating each null hypothesis – from the one associated with the smallest to the one associated with the largest *p*-value – against an α level that is adjusted in order to control for the inflated probability of a Type I error. In this way, the probability of rejecting one or more null hypotheses while they are true will be no larger than 5% (for a proof see Hartley, 1955). Note that for relatively small number of tests *k*, the sequential Bonferroni correction is notably less conservative than the standard Bonferroni correction where one divides α by *k* for all null hypotheses. However, sequential Bonferroni is a conservative procedure in that it always accepts all remaining H_0 whenever one H_0 is not rejected, regardless of how many null hypotheses remain. That is: it does not matter whether one has five or 50 null hypotheses, one single H_0 that is not rejected means that all remaining null hypotheses are also not rejected. As such, some have argued that procedures such as (sequential) Bonferroni, while adequately reducing the probability of a Type I error,

reduce power to find any effect and thus inflate the probability of a Type II error (not rejecting H_0 while the alternative hypothesis H_1 is true; e.g., Benjamini & Yekutieli, 2001; Nakagawa, 2004).

An alternative might be to forego control of FWE and instead control FDR, which is the expected proportion of erroneous rejections of H_0 among all rejections of H_0 (e.g., Benjamini, Drai, Elmer, Kafkaki & Golani, 2001). With the FDR method, the probability of a Type II error is smaller than with sequential Bonferroni but this comes at the expense of a higher probability of a Type I error.

Consequences: Sequential Bonferroni Applied to 60 Published Articles

In our hypothetical example (see **Table 3**) none of the null hypotheses were rejected after sequential Bonferroni correction while, without any correction, all null hypotheses were rejected. One may argue that this example is extreme and contrived, and that such dramatic changes in conclusions will not regularly occur in the empirical literature. We addressed this claim quantitatively, as follows. For each of the six journals at hand, we randomly chose 10 papers per journal that reported the results of one or more multiway ANOVAs. For these 60 papers we re-evaluated their results (see www.aojcramer.com for R-code to perform the sequential Bonferroni procedure; R Development Core Team, 2007) after applying the sequential Bonferroni correction. The results paint a grim picture: in 75% (45/60) of cases, one or more *p*-values were no longer significant. That is, in the majority of cases one or more conclusions are not substantiated by the corrected outcomes of the statistical analyses.

Conclusion

Our literature review showed that, across a total of 819 articles from six leading psychology journals, hardly any researcher corrected for the multiple

comparisons that are an inherent property of multiway ANOVA. A reanalysis of a subset of 60 papers showed that the results of foregoing such correction procedures are worrying: many conclusions reported in the literature may no longer hold after applying a correction procedure. The good news is that sequential Bonferroni (Hartley, 1955) is a simple, easy-to-use correction method that controls the α level, that is, the probability of falsely rejecting true null hypotheses.

One disadvantage of the sequential Bonferroni procedure is conceptual: the significance of a particular factor depends on the significance of other, unrelated factors. For instance, the main effect for G reported in **Table 1** has p = .0329. If the effects for the other two factors (i.e., E*G and E) had been more compelling (e.g., p = .01 for both), the final and third test for G would have been conducted at the α = .05 level, and the result would have been labeled significant. This dependence on the result for unrelated tests may strike one as odd. However, such oddities are a general characteristic of *p*-values (e.g., Wagenmakers, 2007). The regular Bonferroni correction does not have this conceptual drawback, but it is inferior in terms of power.

We do not wish to suggest that the sequential Bonferroni procedure is the only or even the best procedure to correct for multiple comparisons in the multiway ANOVA. As noted before, several other procedures exist. These alternative procedures differ in terms of the balance between safeguarding against Type I and Type II errors. On the one hand, it is crucial to control the probability of rejecting a true null hypothesis (i.e., the Type I error). On the other hand, it is also important to minimize the Type II error, that is, to maximize power (Button et al., 2013)

So what is a researcher to do? Using the sequential Bonferroni correction, one is safeguarded against Type I errors at the expense of failing to detect some effects that are true. Using FDR, one obtains more power, but one relinquishes strict control

over Type I error rate. In such cases, it is sensible to report the results from multiple correction methods: this allows the reader to assess the robustness of the statistical evidence. Of course, the royal road to obtaining sufficient power is not to choose a lenient correction method; instead, one is best advised to increase sample size.

In sum, we have shown that multiway ANOVA harbors a multiplicity problem that has been ignored in empirical practice. The problem can be addressed in a straightforward fashion by the sequential Bonferroni procedure or any other correction procedure. Another fruitful avenue to remedy the problem is *preregistration* (e.g., Chambers, 2013; Chambers, Munafo, et al., 2013; de Groot, 1969; Goldacre, 2009; Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012; Wolfe, 2013). By preregistering their studies and their analysis plan, researchers are forced to consider beforehand the exact hypotheses of interest. In doing so, as we have argued earlier, one engages in confirmative hypothesis testing (i.e., the confirmative multiway ANOVA) in which case one all but solves the multiple comparison problem: The problem disappears when, for example, in a three-way ANOVA one hypothesizes all three implied effects to be present a priori. In that case, in figurative speech, one presses the button in SPSS three separate times, thereby obviating the need for a correction of the alpha level. On the other hand, we argue that “fishing expeditions” in which one has no a priori hypotheses, come at a rather high price: one will have to use some sort of correction procedure to adjust the alpha level when engaging in an explorative multiway ANOVA. This view on differential uses of the multiway ANOVA (i.e., explorative vs. confirmative) hinges on the specific definition of what constitutes a family of hypotheses; and we acknowledge that other definitions of such a family exist. However, in our view, the intentions of the researcher (explorative hypothesis *formation* or confirmative hypothesis *testing*) play a crucial

part in determining the size of the family of hypotheses. It is vital to recognize the multiplicity inherent in the explorative multiway ANOVA and correct the current unfortunate state of affairs[1]; the alternative is to accept that our findings are much less compelling than advertised.

References

Benjamini, Y., Drai, D., Elmer, G., Kafkaki, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research, 125*, 279-284.

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1-12.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610.

Chambers, C. D., Munafo, M., et al. (2013). Trust in science would be improved by study pre-registration. *The Guardian*.

de Groot, A. D. (1969). *Methodology. Foundations of inference and research in the behavioral sciences*. Den Haag, the Netherlands: Mouton.

Didelez, V., Pigeot, I., & Walter, P. (2006). Modifications of the Bonferroni-Holm procedure for a multi-way ANOVA. *Statistical Papers, 47*, 181-209.

[1] Fortunately, some prominent psychologists such as Dorothy Bishop, are acutely aware of the multiple comparison problem in multiway ANOVA and urge their readers to rethink their analysis strategies: http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html.

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

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Goldacre, B. (2009). *Bad science.* London: Fourth Estate.

Hartley, H. O. (1955). Some recent developments in analysis of variance. *Communications on Pure and Applied Mathematics, 8*, 47-72.

Hochberg, Y. (1974). Some generalizations of the *T*-method in simultaneous inference. *Journal of Multivariate Analysis, 4*, 224-234.

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

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217.

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

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

McHugh, R. (1958). Significance level in factorial design. *Journal of Experimental Education, 26*, 257-260.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*, 1044-1045.

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

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage, 40*, 409-414.

Rao, C. R., & Toutenburg, H. (1999). *Linear models: least squares and alternatives*. Springer.

R Development Core Team (2007). R: A language and environment for statistical computing. Version 2.15, URL http://www.R-project.org.

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

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*, 663-665.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40*, 87-104.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81*, 826-831.

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.

Toutenburg, H. (2002). *Statistical analysis of designed experiments.* Springer.

Tukey, J. W. (unpublished). *The problem of multiple comparisons*. Princeton University.

Wagenmakers, E. –J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.

Wagenmakers, E. –J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638.

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics, 75*, 781-783.

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