Bayesian explorations in mathematical psychology

Matzke, D.

Chapter 12

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

A modified version of this chapter has been submitted for publication as:
Angélique O. J. Cramer, Don van Ravenzwaaij, Dora Matzke, Helen Steingroever, Ruud Wetzels, Raoul P. P. P. Grasman, Lourens Waldorp, and Eric-Jan Wagenmakers (2013). Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies.

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%. He we describe the multiple comparison problem and demonstrate that researchers seldom correct for it. We then illustrate the use of the sequential Bonferroni correction —one of several correction procedures— 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 problem in multiway ANOVA is preregistration of the hypotheses.

12.1 Introduction

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 effect of each of the separate factors as well as their interactions. For instance, consider a response time experiment with a 2×3 balanced design (i.e., a design with equal number of participants in each condition); 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, whether factor B is significant, and whether the interaction term $A \times B$ is significant. 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 harbors a multiple comparisons problem, particularly when detailed hypotheses have not been specified a priori (to be discussed in more detail later). Consider, for example, the 2×3 scenario introduced 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. Hence, given the null hypothesis (H_0) is true and $\alpha = 0.05$, the probability of at least one significant result equals $1 - (1 - 0.05)^3 = 0.14$. This is called a Type I error or familywise error rate. The problem of Type I error is not trivial: add a third, balanced factor to the 2×3 scenario (e.g., a $2 \times 3 \times 3$ design), and the probability of finding at least one significant result when H_0 is true increases to $1 - (1 - 0.05)^7 = 0.30$. The situation becomes even more troublesome for designs with unequal numbers of participants per condition: in such unbalanced designs, the three tests in our hypothetical 2×3 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, the α-inflation is dramatic and should be cause for great concern.

The goal of the present article is to highlight the problem of multiple comparison inherent in multiway ANOVA. To this end, we first 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 the pre-registration of the hypotheses of interest.

12.2 Type I Errors and the Oneway ANOVA

A Type I error occurs when the null hypothesis (H_0) is falsely rejected in favor of the 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 $\alpha = 0.05$, the probability of a Type I error is 5%. It is well-known, however, that the multiple comparison problem arises even in the oneway ANOVA whenever the independent variable has more than two levels and post-hoc tests are employed to investigate which condition means differ significantly from one another. Consider, for instance, 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); hence, the researcher will usually perform multiple post-hoc tests to investigate 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 for balanced designs— the overall Type I error equals $1 - (1 - 0.05)^3 = 0.14$. 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 (IBM Corp., 2012) 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, 1973).

12.3 The Explorative Multiway ANOVA: A Family of Hypotheses

Now consider a design that is only slightly more complicated. Suppose a researcher wants to examine the effect of *Ethnicity* (E; three levels) as well as *Gender* (G; two levels) on the total score on a depression questionnaire. Furthermore, suppose that the researcher in question has no firm a priori hypotheses about how E and G influence the depression total score; that is, the researcher is predominantly interested in finding out whether *any* kind of relationship exists between E, G 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 the above example, the multiway ANOVA without strictly formulated a priori hypotheses is an *explorative* one: Using statistical software, such as SPSS, the researcher obtains the results for all three hypotheses involved (i.e., main effect of E, main effect of G, and the $E \times G$ 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*, where "... the term 'family' refers to the collection of hypotheses [...] that is being considered for joint testing" (Lehmann & Romano, 2005). We therefore argue that a multiple comparison problem lurks in the explorative use of the multiway ANOVA.

To see this, consider the results of a fictitious explorative multiway ANOVA reported in Table 12.1. When interpreting the ANOVA table, most researchers would conclude that both main effects as well as the interaction are significant because all p values are smaller than $\alpha = 0.05$. This conclusion is intuitive and directly in line with the numbers reported in Table 12.1. Nevertheless, this conclusion is incorrect. The researcher does not have firm a priori hypotheses and 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 an explorative multiway ANOVA with three factors, we are dealing with seven tests (i.e., three main effects, three first-order interactions, and one second-order interaction), resulting in a 30% Type I error probability; with four factors, the probability of incorrectly rejecting one or more null hypotheses increases to 54%. It is therefore incorrect to evaluate the p values from a multiway ANOVA table with $\alpha = 0.05$.

Note that the above sketched scenario is different from the situation where the researcher uses a multiway ANOVA for *confirmatory* purposes; that is, when 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). For instance, consider a design with two factors and one pre-defined hypothesis: the family is no longer defined as encompassing all three hypotheses implied by the design, but as all to-be-tested hypotheses, in this case a single hypothesis, rendering it unnecessary to adjust the α level.

The realization that explorative multiway ANOVAs inherently harbor a multiple comparison problem may come as a surprise to many empiricists, even to those who regularly use multiway ANOVAs. In standard statistical textbooks, the multiple comparison problem is almost exclusively discussed in the context of oneway ANOVAs. In addition, statistical software packages, such as SPSS, do not present the possible correction procedures for the multiway case, inviting researchers to evaluate the resulting p values with $\alpha = 0.05$.

We are by no means the first to identify the multiplicity problem in multiway ANOVAs (see, e.g., Didelez, Pigeot, & Walter, 2006; Fletcher, Daw, & Young, 1989; Kromrey & Dickinson, 1995;

Table 12.1 Example ANOVA Table for the Three Tests Associated with a 2×3 Design with Gender (G) and Ethnicity (E) as Independent Factors.

Effect	Factor	df_1	df_2	F	p value
Main	G	1	30	5	0.0329*
Main	E	2	30	4	0.0288*
Interaction	$G \times E$	2	30	4.5	0.0195*

Note. *,significant at $\alpha = 0.05$

Olejnik, Li, Supattathum, & Huberty, 1997; Ryan, 1959; R. A. Smith, Levine, Lachlan, & Fediuk, 2002). Earlier work, however, does not feature in mainstream statistical textbooks and is written in a technical style that is inaccessible to scholars without sophisticated statistical knowledge. As a result, empirical work has largely ignored the multiplicity problem. Unfortunately, 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 test only a single pre-specified hypothesis, thereby circumventing the multiple comparison problem. Or perhaps, whenever researchers conduct multiple tests, they use some sort of procedure to adjust the α level for the individual tests. This is unfortunately not the case.

With respect to the former, it is quite common to perform what Gigerenzer (2004) has termed the "null ritual" where researchers specify 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 are often lurked 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 (i.e., Hypothesizing After the Results are Known: HARKing). Hence, hindsight bias and confirmation bias make it difficult for researchers to ignore unexpected "significant" (i.e., individual test with $p < 0.05$) effects.

With respect to the latter, in the next section, we investigate whether researchers correct for multiple comparisons when they use multiway ANOVAs. The short answer is that, almost without exception, researchers interpret the results of the individual tests in isolation, without any correction for multiple comparisons.

12.4 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. Specifically, we investigated all 2010 publications of the following journals:

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).

Table 12.2 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.

Journal	% with mANOVA	% with mANOVA & correction
Journal of Experimental Psychology: General	84.61	0
Psychological Science	43.16	0
Journal of Abnormal Psychology	31.82	0
Journal of Consulting and Clinical Psychology	16.30	0
Journal of Experimental Social Psychology	65.17	2.59
Journal of Personality and Social Psychology	54.41	1.35
Overall	47.62	1.03

Note. Overall, all papers from the six journals; mANOVA, multiway ANOVA.

6. *Journal of Personality and Social Psychology* (volumes 98-99; issues 1-6; 136 papers).

For each article, we assessed whether the papers featured one of more multiway ANOVAs and whether the authors had reported some sort of correction procedure (e.g., an omnibus test; see below) to remedy the multiple comparison problem. The results are summarized in Table 12.2.

Two results stand out. First, almost half of the articles under investigation here used a multiway ANOVA, underscoring the popularity of this testing procedure. Second, only around 1% of the articles used a correction procedure. In all four cases where a correction procedure was used, this was an omnibus F test, where 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 H0 is rejected as well as all other combinations of null hypotheses (Fletcher et al., 1989; Wright, 1992). For example, for a 2×2 ANOVA, one should first test the omnibus hypothesis with all three hypotheses included (two main effects and the 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 for 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 affairs 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).

12.5 Possible Remedy: Sequential Bonferroni Correction

As noted earlier, statisticians have long been aware of the multiple comparison problem in multiway ANOVAs. 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 threeway ANOVA, a main effect is present

Table 12.3 The Sequential Bonferroni Procedure for the Hypothetical Example of Table 1.

Effect	p value	α_{adj}	H_0
$G \times E$	0.0195	0.0167	not rejected
E	0.0288	0.0250	not rejected
G	0.0329	0.0500	not rejected

Note. The sequential Bonferroni procedure entails: (1) sorting p values in ascending order; (2) computing adjusted α level per test (α_{adj}); (3) sequentially evaluating each p value with adjusted α level (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).

for one factor but not for the remaining two factors; then the overall F test is likely to yield a significant F value because, indeed, the omnibus H_0 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). The sequential Bonferroni correction 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 twoway ANOVA with G and E as independent factors (uncorrected results are listed in Table 12.1). The results of the sequential Bonferroni correction procedure for this example are presented in Table 12.3. 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. In the present example, we conduct three tests so α_{adj} for the smallest p value equals $0.05/3 = 0.01667$. For the second p value, α_{adj} equals α divided by the number of tests minus 1: $\alpha_{adj} = 0.05/2 = 0.025$. For the final p value, α_{adj} equals α divided by the total number of tests minus 2: $\alpha_{adj} = 0.05/1 = 0.05$. Next, one evaluates each p value with the adjusted α level, sequentially, with the smallest p value evaluated first. Importantly, if the H_0 associated with a given p value is not rejected (i.e., $p > \alpha_{adj}$), all testing ends and all remaining tests are also considered non-significant.

In our example, we evaluate $p = 0.0195$ with $\alpha_{adj} = 0.01667$: $p > \alpha_{adj}$ and therefore we conclude that the $G \times E$ interaction is not significant. We therefore stop testing and conclude that the remaining main effects are not significant either. Thus, with the sequential Bonferroni correction procedure, we conclude that none of the effects are significant; without the correction procedure, we had concluded that all of the effects were significant.

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

Thus, the sequential Bonferroni correction procedure allows one to control for the familywise error by evaluating each H_0 –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 incorrectly rejecting one or more null hypotheses will be no larger than 5% (for a proof, see Hartley, 1955). Note that for a relatively small number of tests k, the sequential Bonferroni correction is notably less conservative than the standard Bon-

ferroni correction where one divides α by k for all null hypotheses. However, sequential Bonferroni is a conservative procedure in that it never rejects the remaining null hypotheses whenever a given H_0 is not rejected, regardless of how many null hypotheses remain. That is, it does not matter whether we deal with five or 50 null hypotheses, one single H_0 that is not rejected means that the remaining null hypotheses cannot be rejected either. As such, it has been argued that procedures such as (sequential) Bonferroni, while adequately reducing the probability of a Type I error, reduce power and hence inflate the probability of a Type II error (i.e., not rejecting H_0 when H_1 is true; e.g., Benjamini & Yekutieli, 2001; Nakagawa, 2004).

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

12.6 Consequences: Sequential Bonferroni Applied to 60 Published Articles

In our hypothetical example, none of the null hypotheses were rejected after the sequential Bonferroni correction (see Table 12.3), whereas, without any correction, all null hypotheses were rejected (see Table 12.1). 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 listed in Table 12.2, we randomly chose 10 articles that reported one or more multiway ANOVAs. For these 60 papers, we re-evaluated the results (see `www.aojcramer.com` for R code (R Core Team, 2012) to perform the sequential Bonferroni procedure) after applying the sequential Bonferroni correction. The results paint a grim picture: in 75% (45/60) of the cases, one or more p values were no longer significant after correcting for multiple comparisons. That is, in the majority of cases, one or more conclusions are not substantiated by the corrected outcomes of the statistical analyses.

12.7 Conclusion

Our literature review showed that, across a total of 819 articles from six leading psychology journals, hardly any researchers have corrected for the multiple comparison problem that is 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 the sequential Bonferroni procedure (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 12.1 and Table 12.3 is associated with $p = 0.0329$. If the effects for the other two factors (i.e., $E \times G$ and E) had been more compelling (e.g., $p < 0.01$), the final and third test for G would have been conducted with $\alpha = 0.05$ level, and the result would have been labeled significant. This dependence on the results of unrelated tests may strike one as odd. However, such oddities are a general characteristic of p values (e.g., Wagenmakers, 2007). Note that 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 incorrectly rejecting the H_0 (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 the false discovery rate procedure, one obtains more power, but relinquishes strict control over the Type I error rate. We encourage researchers to report the results from multiple correction methods: this allows readers to assess the robustness of the statistical evidence. Of course, the royal road to obtaining sufficient power is not to choose lenient correction methods; 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 various correction procedures, such as the sequential Bonferroni correction. Another fruitful avenue to remedy the problem is the *pre-registration* of the hypotheses and the corresponding statistical analyses (e.g., Chambers, 2013; Chambers et al., 2013; de Groot, 1969; Goldacre, 2009; Wagenmakers et al., 2012; Wolfe, 2013). Pre-registration forces researchers 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). "Fishing expeditions", however, 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 α level when engaging in an explorative multiway ANOVA.

The 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* vs. 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.

[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`.