

Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies

Angélique O. J. Cramer¹ · Don van Ravenzwaaij² · Dora Matzke¹ · Helen Steingroever¹ · Ruud Wetzels³ · Raoul P. P. Grasman¹ · Lourens J. Waldorp¹ · Eric-Jan Wagenmakers¹

Published online: 15 September 2015

© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract Many psychologists do not realize that exploratory use of the popular multiway analysis of variance 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 at least one Type I error (if all null hypotheses are true) is 14 % rather than 5 %, if the three tests are independent. We explain the multiple-comparison problem and demonstrate that researchers almost never correct for it. To mitigate the problem, we describe four remedies: the omnibus F test, control of the familywise error rate, control of the false discovery rate, and preregistration of the hypotheses.

Keywords Multiway ANOVA · Factorial ANOVA · Multiple comparison problem · Type I error · Sequential Bonferroni ·

Benjamini–Hochberg procedure · Familywise error rate · False discovery rate · 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 2×3 balanced design (i.e., a design with equal numbers 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 $A \times B$ is significant (at the .05 level). In the same vein, the standard multiway ANOVA is also frequently used in nonexperimental 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 the problem of multiple comparisons, in particular when detailed hypotheses have not been specified a priori (to be discussed in more detail later). For the 2×3 scenario discussed above, without a priori hypotheses (i.e., when the researcher’s attitude can be best described by “let us see what we can find”; de Groot, 1969), the probability of

✉ Angélique O. J. Cramer
aoj.cramer@gmail.com

¹ Psychological Methods, Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

² Faculty of Science and Information Technology, School of Psychology, University of Newcastle, Callaghan, New South Wales, Australia

³ Data Analytics, Price Waterhouse Coopers, Amsterdam, The Netherlands

finding at least one significant result, given that the data in fact originate from the null hypotheses, lies in the vicinity of $1 - (1 - .05)^3 = .14$.¹ This is called the *Type I error* or *familywise error rate*. The problem of Type I errors 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 the null hypothesis is true increases to around 30 % [$1 - (1 - .05)^7$], the precise probability depending on the extent to which the tests are correlated (see also note 1). Thus, in the absence of strong a priori expectations about the tests that are relevant, this alpha inflation can be substantial and be a cause for concern.

Here we underscore the problem of multiple comparisons inherent in the exploratory multiway ANOVA. We conduct a literature review and demonstrate that the problem is widely ignored: Recent articles published in six leading psychology journals contain virtually no procedures to correct for the multiple-comparison problem. Next, we outline four possible remedies: the omnibus *F* test, control of the family wise error rate using the sequential Bonferroni procedure, control of the false discovery rate using the Benjamini–Hochberg procedure, and the preregistration of hypotheses.

Background: Type I errors and the one-way 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 one-way ANOVA, the probability of a Type I error can be controlled by setting the significance level α . For example, when $\alpha = .05$, the probability of a Type I error is 5 %. Since the one-way ANOVA comprises only one test, there is no multiple-comparison problem. It is well-known, however, that this problem arises in the one-way 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 one-way ANOVA and obtains a significant effect of ethnicity on the total score of a depression questionnaire. Assume that

ethnicity has three levels (e.g., Caucasian, African-American, and Asian); this researcher will then usually perform multiple post-hoc tests to determine which ethnic groups differ significantly from one another. Here the three post-hoc tests are Caucasian versus African-American, Caucasian versus Asian, and African-American versus Asian. Fortunately, for the one-way ANOVA the multiple-comparison problem has been thoroughly studied. Software programs such as SPSS and SAS explicitly address 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, 1994; Westfall, Tobias, & Wolfinger, 2011).

The exploratory 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 *exploratory* one: Without strictly formulated a priori hypotheses, the researcher obtains the results for all three hypotheses involved (i.e., main effect of G, main effect of E, and a $G \times E$ interaction) by means of a single mouse click in SPSS. As such, in an exploratory 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 exploratory uses of a multiway ANOVA.

To see this, consider the results of a fictitious exploratory multiway ANOVA, reported in Table 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 $\alpha = .05$. This conclusion is intuitive and directly in line with the numbers reported in Table 1. Nevertheless, this conclusion is statistically unwarranted; the researcher does not have firm a priori hypotheses and therefore tests all three hypotheses simultaneously, engaging in an exploratory research effort. In this case, when all null hypotheses are true, the Type I error will be larger than 5 % (around

¹ The probability of finding at least one significant result equals exactly 14 % if and only if the three tests are completely independent. This is only true if the total number of participants in the sample approaches infinity: In that case, the *F* tests become asymptotically independent. For all other sample sizes, the test statistics are not independent, because they share a common value—namely the mean square error in the denominator (Feingold & Korsog, 1986; Westfall, Tobias, & Wolfinger, 2011). This induces dependence among the test statistics. Another way in which dependence between the tests is induced is when the design is *unbalanced*—that is, with unequal numbers of participants per condition. The consequence of the dependence between the test statistics is that the probability of finding at least one significant result, given that all null hypotheses are true, will be slightly lower than 14 %.

Table 1 Example ANOVA table for the three tests associated with a hypothetical 2×3 design with Gender (G) and Ethnicity (E) as independent factors

		<i>df</i> 1	<i>df</i> 2	<i>F</i>	<i>p</i> value
Main effect	G	1	30	5	.0329*
	E	2	30	4	.0288*
Interaction	G \times E	2	30	4.50	.0195*

* significant at $\alpha = .05$

14 %—see note 1). 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, without a priori hypotheses and when all null hypotheses are true, 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 around 30 % [i.e., $1 - (1 - .05)^7$]; with four factors and when all null hypotheses are true, the probability of incorrectly rejecting one or more null hypotheses is around 54 %. It is therefore incorrect to compare each of the *p* values from a multiway ANOVA table to $\alpha = .05$.

This is notably different from the situation in which the researcher uses a multiway ANOVA for *confirmatory* purposes; that is, 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 .

The realization that exploratory 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 (with Westfall, Tobias, & Wolfinger, 2011, as a notable exception). 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 $\alpha = .05$.

We are not the first to identify the multiplicity problem in the multiway ANOVA (see, e.g., Didelez, Pigeot, & Walter, 2006; Fletcher, Daw, & Young, 1989; Kromrey & Dickinson, 1995; Olejnik, Li, & Supattatum, 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 in the multiway ANOVA. 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 prespecified 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 the 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 (i.e., hypothesizing after the results are known, or HARKing; see also Barber, 1976). 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 $p < .05$).

The next section addresses 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.

Prevalence: Multiway corrections in six psychology 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 articles)
2. *Psychological Science*, volume 21, issues 1–12 (285 articles)
3. *Journal of Abnormal Psychology*, volume 119, issues 1–4 (88 articles)
4. *Journal of Consulting and Clinical Psychology*, volume 78, issues 1–6 (92 articles)
5. *Journal of Experimental Social Psychology*, volume 46, issues 1–6 (178 articles)

6. *Journal of Personality and Social Psychology*, volumes 98 and 99, issues 1–6 (136 articles)

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.

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 studies used a correction procedure (i.e., the omnibus F test; see below).

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, in which the problem is more evident and it is common practice to correct for multiplicity (e.g., Poldrack et al., 2008).

Remedies

As we 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. Below we discuss four different procedures to mitigate the multiple-comparison problem in multiway ANOVA: (1) the omnibus F test, (2) controlling the family

wise error rate, (3) controlling the false discovery rate, and (4) preregistration.

Remedy 1: The omnibus F test

In the few cases in which a correction procedure was used, this involved 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 this omnibus H_0 is rejected (Fletcher, Daw, & Young, 1989; Wright, 1992). So, for example, in the case of a 2×2 ANOVA, one should first test the omnibus hypothesis with all three hypotheses included (two main effects and an interaction). If this result is significant, one may continue and test the individual hypotheses.

However, the omnibus F test does not 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. Hence, the omnibus F test offers only weak protection against the multiplicity problem.

Remedy 2: Controlling familywise error rate

The familywise error rate (FWER) refers to the probability of making at least one Type I error within the family of tests under consideration; here, the family consists of all tested effects in a multiway ANOVA without a priori hypotheses. To control this FWER, one has to make certain that it is smaller than or equal to α , which usually equals 5 %. Preferably, FWER is controlled in the *strong sense*, such that it holds for any configuration of true and false null hypotheses.

One method to control FWER in the strong sense is the sequential Bonferroni procedure (also known as the *Bonferroni–Holm correction*), which was first introduced by Hartley (1955) and subsequently (independently) reinvented and/or modified by others (Hochberg, 1988; Holm, 1979; McHugh, 1958; Rom, 1990; Shaffer, 1986; Wright, 1992). To illustrate the procedure, let us revisit our hypothetical example in which a researcher conducts a two-way ANOVA with G and E as independent factors (the 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 (see also Fig. 1 for a visual explanation of the method). 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

Table 2 Percentages of articles overall and in the six selected journals that used a multiway analysis of variance (mANOVA), and the percentages of these articles that used some sort of correction procedure

	% Articles Using mANOVA	% Articles Using mANOVA + Correction
Overall	47.62	1.03
<i>JEPG</i>	84.61	0
<i>Psych Sci</i>	43.16	0
<i>J Abn Psych</i>	31.82	0
<i>JCCP</i>	16.30	0
<i>JESP</i>	65.17	2.59
<i>JPSP</i>	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*

.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 nonsignificant as well.

In our example, we evaluate $p = .0195$ against $\alpha_{adj} = .01667$: $p > \alpha_{adj}$, and therefore we conclude that the $G \times E$ interaction is not significant. The sequential Bonferroni procedure mandates that we stop testing, and we conclude that the remaining main effects are not significant, either. Thus, when the sequential Bonferroni correction procedure is applied to our example, none of the effects are significant; without a correction procedure, all of the effects are significant.

Thus, the sequential Bonferroni correction procedure allows control over the FWER 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 numbers of tests k , the sequential Bonferroni correction is notably less conservative than the standard Bonferroni correction, in which one divides α by k for all null hypotheses. However, sequential Bonferroni is still a relatively conservative procedure, in that it always retains the remaining H_0 s whenever one H_0 is not rejected, regardless of how many 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 controlling 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 when the

alternative hypothesis H_1 is true; see, e.g., Benjamini & Yekutieli, 2001; Nakagawa, 2004).

Another 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., $G \times E$ 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 $\alpha = .05$ level, and the result would have been labeled significant. This dependence on the results from unrelated tests may strike one as odd.

The sequential Bonferroni procedure is by no means the only one in its class, and we present it here merely as a prototypical example of a procedure that seeks to control FWER. A well-known alternative procedure is the regular Bonferroni

Table 3 Results from the sequential Bonferroni (seqB) and Benjamini–Hochberg (BH) procedures for the example from Table 1

Effect	p Value	α_{adj} seqB	α_{adj} BH	H_0 seqB	H_0 BH
$G \times E$.0195	.0167	.0167	retained	rejected
E	.0288	.0250	.0333	retained	rejected
G	.0329	.0500	.0500	retained	rejected

α_{adj} seqB = the adjusted alpha level with the sequential Bonferroni procedure; α_{adj} BH = the adjusted alpha level with the Benjamini–Hochberg procedure; H_0 seqB = evaluation of the null hypotheses with the sequential Bonferroni procedure; H_0 BH = evaluation of the null hypotheses with the Benjamini–Hochberg procedure.

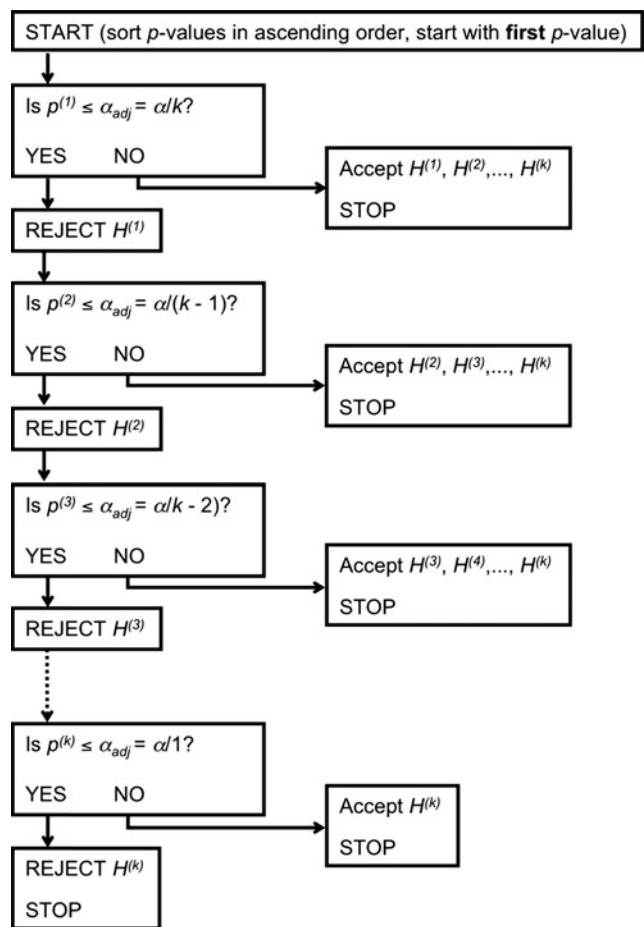


Fig. 1 A visual representation of the sequential Bonferroni method for controlling familywise error rate. All p values are sorted in ascending order and are assigned a rank number from 1 (smallest) to k (largest). Next, one starts by evaluating the first (smallest) p value ($p^{(1)}$) against the adjusted α (α_{adj}), which is—for the first p value—equal to α divided by k . If the p value is smaller than α_{adj} , then the first hypotheses $H^{(1)}$ is rejected, and one proceeds to the second p value. If the p value is not smaller than α_{adj} , then one immediately accepts all null hypotheses and stops testing

correction, in which α , for every p value alike, is divided by the total number of tests. As such, the regular Bonferroni correction does not have the conceptual drawback of the significance of one result being dependent on the other results for unrelated tests. However, as compared to sequential Bonferroni, the regular Bonferroni is inferior in terms of power. Other methods to control FWER are, for example, the Simes procedure (Simes, 1986) and the Hommel correction (Hommel, 1988).

Remedy 3: Controlling false discovery rate

An alternative might be to forgo control of FWER and instead to control the *false discovery rate* (FDR; Benjamini, Drai, Elmer, Kafkafi, & Golani, 2001; Benjamini & Hochberg, 1995), which is the expected proportion of erroneous rejections of H_0 among all rejections of H_0 . When controlling FDR, the probability of a Type II error is smaller than when controlling FWER, but this comes at the expense of a higher probability of a Type I error. Controlling FDR is particularly appropriate for applications in genetics and neuroimaging, in which the goal is to identify candidate effects from a large set; these candidates can then be tested more rigorously in follow-up confirmatory experiments.

One way to control FDR is with the Benjamini–Hochberg procedure (BH; Benjamini & Hochberg, 1995). To illustrate the procedure, consider again our hypothetical example for which the uncorrected results are listed in Table 1. The results of the BH procedure for this example are presented in Table 3, and they were obtained as follows: First, one sorts all p values in ascending order—that is, with the smallest p value first (see also Fig. 2 for a visualization of the method). Next, one computes an adjusted α level, α_{adj} . For the largest p value, α_{adj} equals α times the rank number of the largest p value (3, in our example), divided by the total number of tests (also 3 in this example): $.05 \times (3/3) = .05$. For the middle p value, α_{adj} equals $.05 \times (2/3) = .0333$; for the smallest p value, α_{adj} equals $.05 \times (1/3) = .01667$. Next, one evaluates each p value against these adjusted α levels, with the largest p value being evaluated *first*. Importantly, if the H_0 associated with this p value is rejected (i.e., $p < \alpha_{\text{adj}}$), then all testing ends and all remaining tests are considered significant as well.

In our example, we evaluate $p = .0329$ against $\alpha_{\text{adj}} = .05$: $p < \alpha_{\text{adj}}$, and therefore we conclude that the main effect of G is significant (and thus, H_0 is rejected). According to the BH procedure, we stop testing and conclude that this main effect, the other main effect, and the interaction are all significant. Note that this conclusion is drawn despite the fact that the p value for the $G \times E$ interaction exceeded the adjusted alpha level. In the alternative situation, that we would have retained the null

hypothesis of the first p value, the testing would have continued by evaluating the second p value against its adjusted alpha.

The BH procedure is certainly not the only way to control FDR. Other procedures include the Benjamini–Hochberg–Yekutieli procedure (Benjamini & Yekutieli, 2001), which controls FDR under positive dependence assumptions, and the Efron method (Efron, Storey, & Tibshirani, 2001; Efron, Tibshirani, Storey, & Tusher, 2001), which controls not exactly FDR, but *local* FDR, which is the conditional probability that the null hypothesis is true given the data.

Remedy 4: Preregistration

Another effective remedy is *preregistration* (e.g., Chambers, 2013; Chambers et al., 2013; de Groot, 1969; Goldacre, 2009; Nosek & Lakens, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012; Wolfe, 2013; for preregistration in medical clinical trials, see, e.g., www.clinicaltrials.gov). By preregistering their studies and their analysis plans, researchers are forced to specify beforehand the exact hypotheses of interest. In doing so, as we have argued earlier, one engages in confirmatory hypothesis testing (i.e., the confirmatory multiway ANOVA), a procedure that can greatly mitigate the multiple-comparison problem. For instance, consider experimental data analyzed with a $2 \times 2 \times 3$ multiway ANOVA; if the researcher stipulates in advance that the interest lies in the three-way interaction and the main effect of the first factor, this reduces the number of tested hypotheses from seven to two, thereby diminishing the multiplicity concern.

Conclusion

We have argued that the multiway ANOVA harbors a multiple-comparison problem, particularly when this analysis technique is employed relatively blindly—that is, in the absence of strong a priori hypotheses. Although this hidden multiple-comparison problem has been studied in statistics, empiricists are not generally aware of the issue. This point is underscored by our literature review, which showed that, across a total of 819 articles from six leading journals in psychology, corrections for multiplicity are virtually absent.

The good news is that the problem, once acknowledged, can be remedied in one of several ways. For instance, one could use one of several procedures to control either the FWER (e.g., with the sequential Bonferroni procedure) or the FDR (e.g., with the Benjamini–Hochberg procedure). These procedures differ in terms of the balance between safeguarding against Type I and Type II errors. On the one

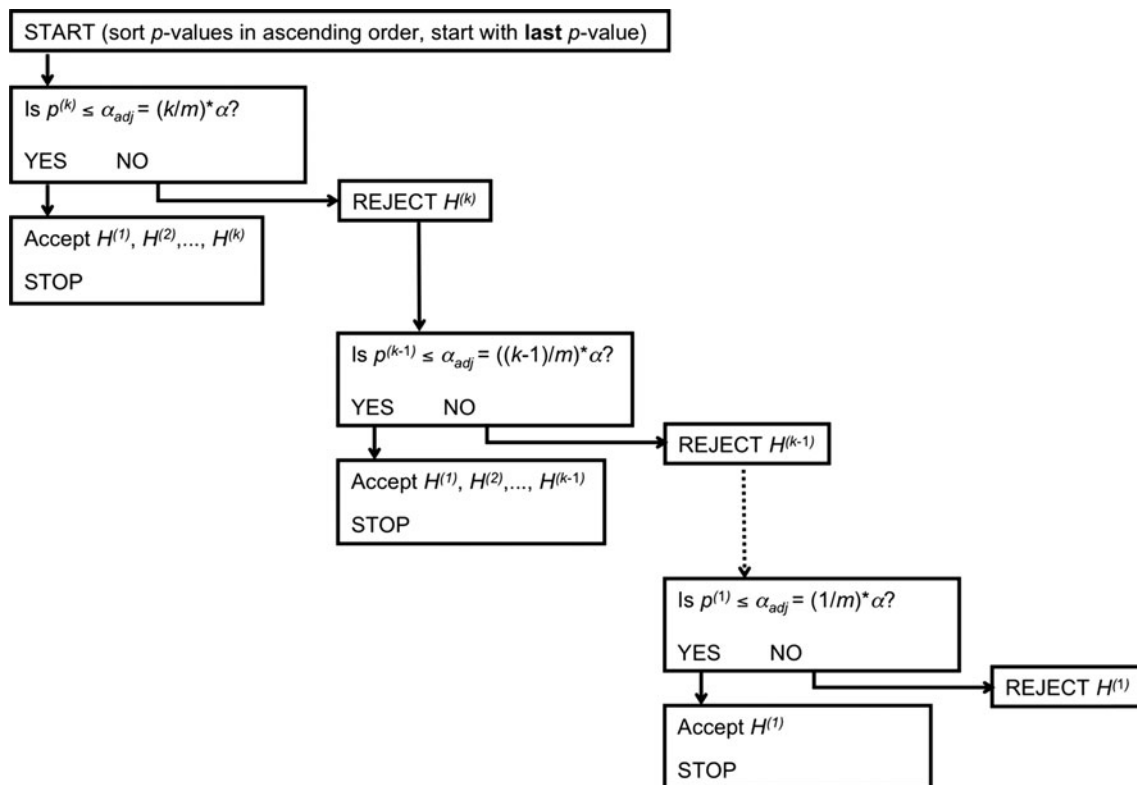


Fig. 2 A visual representation of the Benjamini–Hochberg procedure for controlling false discovery rate. All m p values are sorted in ascending order and assigned a rank number from 1 (smallest) to k (largest). Next, one starts by evaluating the last (largest) p value ($p^{(k)}$) against the

adjusted α (α_{adj}), which is—for the last p value—equal to k divided by m times α . If the p value is smaller than α_{adj} , then all null hypotheses are rejected and testing stops. If the p value is not smaller than α_{adj} , then one proceeds to the next p value

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). As we have shown in our fictitious data example, the side toward which the balance shifts may make a dramatic difference in what one would conclude from the data: When using sequential Bonferroni (i.e., better safeguard against Type I errors at the cost of a reduction in power), all null hypotheses were retained; when using the Benjamini–Hochberg procedure (i.e., less control over Type I errors but more power), all null hypotheses were rejected. So, what is a researcher to do when various correction procedures result in such different conclusions? It appears prudent to follow the statistical rule of thumb for handling uncertainty: When in doubt, issue a full report that includes the results from all multiple-correction methods that were applied. Such a full report 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 plan for a large sample size (Klugkist, Post, Haahruijs, & van Wesel, 2014).

And there is even better news. Many, if not all, correction methods for controlling either FWER or FDR are easy to implement using the function `p.adjust()` in the basic *stats*

package in R (R Development Core Team, 2007). All that is required is to input a vector of p values, and the function evaluates these according to the chosen correction method.

We realize that our view on differential uses of the multiway ANOVA (i.e., exploratory vs. confirmatory) 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 (exploratory hypothesis *formation* or confirmatory 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 exploratory multiway ANOVA and to correct the current unfortunate state of affairs²; the alternative is to accept that our findings might be less compelling than advertised.

² 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>.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. New York, NY: Pergamon Press.
- Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, *125*, 279–284. doi:10.1016/S0166-4328(01)00297-2
- 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. *Annals of Statistics*, *29*, 1165–1188.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. doi:10.1038/nrn3475
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610. doi:10.1016/j.cortex.2012.12.016
- Chambers, C. D., Munafò, M., et al. (2013, 5 June). Trust in science would be improved by study pre-registration. *The Guardian*. Retrieved from www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration
- de Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences*. The Hague, 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.
- Efron, B., Storey, J., & Tibshirani, R. (2001). *Microarrays, empirical Bayes methods, and false discovery rates* (Technical Report, July 2001). Stanford, CA: Stanford University, Department of Statistics. Retrieved from <http://statweb.stanford.edu/~ckirby/brad/papers/2001MicroEBMethods.pdf>
- Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001b). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*, 1151–1160.
- Feingold, M., & Korsog, P. E. (1986). The correlation and dependence between two F statistics with the same denominator. *American Statistician*, *40*, 218–220.
- 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. *Journal of Socio-Economics*, *33*, 587–606.
- Goldacre, B. (2009). *Bad science*. London, UK: 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. Retrieved from www.jstor.org/stable/4615733
- Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*, 383–386.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.
- Klugkist, I., Post, L., Haahruijs, F., & van Wesel, F. (2014). Confirmatory methods, or huge samples, are required to obtain power for the evaluation of theories. *Open Journal of Statistics*, *4*, 710–725.
- 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. *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.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.
- Olejnik, S., Li, J., & Supattatham, 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.
- R Development Core Team. (2007). *R: A language and environment for statistical computing (Version 2.15)*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org
- Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*, 663–665.
- Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, *56*, 26–47. doi:10.1037/h0042478
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, *40*, 87–110. doi:10.1093/biomet/40.1-2.87
- Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826–831.
- Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754.
- 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.
- Tukey, J. W. (1994). The problem of multiple comparisons. In H. I. Braun (Ed.), *The collected works of John W. Tukey: Vol. 8. Multiple comparisons: 1948–1983* (pp. 1–300). New York, NY: Chapman and Hall.
- 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. doi:10.1177/1745691612463078
- Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS* (2nd ed.). Cary, NC: SAS Institute Inc.
- Wolfe, J. M. (2013). Registered reports and replications in *Attention, Perception, & Psychophysics* [Editorial]. *Attention, Perception, & Psychophysics*, *75*, 781–783. doi:10.3758/s13414-013-0502-5
- Wright, S. P. (1992). Adjusted p -values for simultaneous inference. *Biometrics*, *48*, 1005–1013.