Do *p* Values Lose Their Meaning in Exploratory Analyses? It Depends How You Define the Familywise Error Rate

Mark Rubin
The University of Newcastle, Australia

Citation: Rubin, M. (2017). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, *21*, 269-275. doi: 10.1037/gpr0000123

Abstract

Several researchers have recently argued that *p* values lose their meaning in exploratory analyses due to an unknown inflation of the alpha level (e.g., Nosek & Lakens, 2014; Wagenmakers, 2016). For this argument to be tenable, the familywise error rate must be defined in relation to the number of hypotheses that are tested in the same study or article. Under this conceptualization, the familywise error rate is usually unknowable in exploratory analyses because it is usually unclear how many hypotheses have been tested on a spontaneous basis and then omitted from the final research report. In the present article, I argue that it is inappropriate to conceptualize the familywise error rate in relation to the number of hypotheses that are tested. Instead, it is more appropriate to conceptualize familywise error in relation to the number of different tests that are conducted on the same null hypothesis in the same study. Under this conceptualization, alpha level adjustments in exploratory analyses are (a) less necessary and (b) objectively verifiable. As a result, *p* values do not lose their meaning in exploratory analyses.

Keywords: confirmatory research; exploratory research; familywise error rate; *p* values; Type I errors

Correspondence concerning this article should be addressed to Mark Rubin at the School of Psychology, Behavioural Sciences Building, The University of Newcastle, Callaghan, NSW 2308, Australia. Tel: +61 (0)2 4921 6706. Fax: +61 (0)2 4921 6980. E-mail: Mark.Rubin@newcastle.edu.au Web: http://bit.ly/QgpV4O

The replication crisis (e.g., Munafò et al., 2017) has led several researchers to conclude that it is inappropriate to interpret *p* values in exploratory analyses (Nosek, Ebersole, DeHaven, & Mellor, 2017; Nosek & Lakens, 2014; Forstmeier, Wagenmakers, & Parker, 2016; Wagenmakers, 2016; see also Dahl, Grotle, Benth, & Natvig, 2008; De Groot, 1956/2014). These researchers argue that "in exploratory analysis, *p*-values lose their meaning due to an unknown inflation of the alpha-level" (Nosek & Lakens, 2014, p. 138). Wagenmakers (2016) has recently used this argument to support the case for preregistered research:

> Statistical tools such as *p*-values and confidence intervals are meaningful only for strictly confirmatory analyses. In turn, preregistration is one of very few ways to check and confirm that the presented analyses were indeed confirmatory. Two conclusions follow:
>
> (1) Researchers who do exploratory work cannot interpret the outcome of their statistical tests in a meaningful way. The problem is one of multiple comparisons with the number of comparisons unknown (De Groot, 1956/2014) — in other words, cherry-picking.
>
> (2) Researchers who wish to interpret the outcome of their statistical tests in a meaningful way are forced to preregister their analyses. Preregistration is the price one pays for being allowed anywhere near a statistical test.

If these conclusions are correct, then the implications are far-reaching. They suggest that a substantial amount of non-preregistered research that has used *p* values in null hypothesis significance tests may be invalid.

In this article, I rebut the argument that *p* values lose their meaning in exploratory analyses. I agree that it is difficult to compute alpha level inflation in exploratory analyses if that computation is based on the number tests that are used to investigate *all* of the hypotheses that are tested in a study. However, I argue that it is inappropriate to compute alpha level inflation in this way. Instead, following Matsunaga (2007), I argue that it is more appropriate to compute alpha inflation based on the number of different tests that are used to investigate *single* null hypotheses. Under this conceptualization, (a) alpha level adjustment is less necessary in exploratory analyses, and (b) when it is necessary, it is localized to specific reported hypotheses, making it "knowable" (i.e., objectively verifiable). Based on these points, I argue that *p* values do not lose their meaning in exploratory analyses.

To be clear, nothing in my argument is intended to diminish either the seriousness of the replication crisis or the usefulness of preregistering research protocols. Furthermore, I acknowledge that it is important for researchers to distinguish between confirmatory tests and exploratory investigations in their research reports (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). My intention is only to put forward a counterargument to the proposition that *p* values necessarily lose their meaning in exploratory analyses. I begin by explaining how multiple testing causes alpha inflation.

How does Multiple Testing Cause Alpha Inflation?

The *Type I error rate* (α) is the probability of incorrectly rejecting a null hypothesis in a long run of exact replications in which samples are randomly drawn from the exact same population.[1] Type I errors occur because the random sampling procedure occasionally produces a sample that is unrepresentative of the specified population (i.e., sampling error). Researchers usually prespecify the Type I error rate as a *nominal alpha level*. For example, if a researcher sets their nominal alpha level at .05, then it means that they are willing to accept a 5% probability that any result that they categorize as "significant" based on their observed *p* values is a spurious result that occurs due to random sampling error in the long run.

The Type I error rate, or *actual alpha level*, increases as a function of the number of tests that are conducted. For example, if a researcher conducts a single test with a nominal alpha level of .05, then they have a 5% probability of making a Type I error in the long run. However, if they conduct two such tests, and they reject the null hypothesis if either test yields a significant result, then the combined probability of making a Type I error across the two tests increases to $1 - (1 - .05)^2 = 9.75\%$. Hence, their actual alpha level becomes inflated relative to their prespecified nomimal alpha level, meaning that they have a greater probability of incorrectly rejecting the null hypothesis than they originally specified. This *inflated alpha level* occurs because the researcher's nominal alpha level was only intended to cover *one* opportunity to incorrectly reject the null hypothesis but it is actually applied to *two* opportunities (i.e., two tests rather than one).

The combined Type I error rate for multiple tests is called the *familywise error rate* because it refers to the probability of making at least one Type I error among a collection or *family* of tests. In order to maintain a prespecified nominal alpha level of .05 across a family of tests, researchers need to reduce the nominal alpha level for each individual test as a function of the number of tests that are included in a family. In the previous example, the researcher would need to reduce their nominal alpha level for each of the two tests from .05 to .025 in order to maintain an overall nominal alpha level of .05 for the null hypothesis that they are testing.

Why is Multiple Testing Problematic in Exploratory Analyses?

In order to adjust the nominal alpha level to take account of the familywise error rate, the precise number of tests in the family needs to be known. The problem that has been pointed out by Nosek, Wagenmakers, and others (Bender & Lange, 2001; Dahl et al., 2008; De Groot, 1956/2014; Forstmeier et al., 2016; Nosek et al., 2017; Nosek & Lakens, 2014; Wagenmakers, 2016) is that although the precise number of tests is known in preregistered confirmatory analyses, it is not usually known in exploratory analyses. This is because exploratory analyses tend to involve a lot of tests, and only a subset of those tests is documented. Wagenmakers and colleagues (in De Groot, 1956/2014, p. 193) explained De Groot's description of this problem as follows:

> De Groot suggests that the effect of exploration is similar to that of conducting multiple tests, except that in exploration the number of tests conducted or intended remains undetermined—consequently, it is impossible to correct statistically for the fact that a test was exploratory rather than confirmatory.

Similarly, Forstmeier et al. (2016, p. 10) explained that:

> the main problem with data exploration is that we normally do not keep track of the number of tests that we have conducted or would have been willing to entertain, so there exists no objective way of correcting for the extent of multiple testing (De Groot, 1956/2014).

I argue that keeping track of the number of tests is only problematic for alpha level adjustment when researchers adopt a specific approach to the familywise error rate that is based on *tests of multiple hypotheses*. I explain why this multiple hypotheses approach is inappropriate, and I then discuss a second approach that is based on *different tests of the same hypothesis*. I demonstrate that when this second approach is used, the precise number of tests that are conducted can be more easily determined and, consequently, it is possible to correct for alpha level inflation, even in exploratory analyses.

Two Approaches to the Familywise Error Rate

Familywise Error Rates Based on Tests of Multiple Hypotheses

The argument that *p* values lose their meaning in exploratory analyses assumes that familywise error rates are based on tests of *multiple hypotheses*. Under this operationalization, the familywise error rate relates to all of the tests of a collection of *several different hypotheses*. For example, according to Cramer, Wagenmakers and colleagues: "the term 'family' refers to the *collection of hypotheses* ...that is being considered for joint testing" (Lehmann & Romano, 2005, as cited in Cramer et al., 2016, p. 641, my emphasis). If the familywise error rate is operationalized in this way, then it is correct to conclude that *p* values lose their meaning in exploratory analyses because the number of hypotheses that is tested is undocumented and unknowable and, therefore, the alpha level inflation cannot be estimated. In addition, as Wagenmakers (2016) pointed out, this problem does not affect confirmatory analyses because in this case the number of hypotheses is predetermined and recorded in a preregistered research protocol. Consequently, although alpha level inflation is unknowable and therefore uncorrectable in exploratory analyses, it is knowable and correctable in preregistered confirmatory analyses.

However, operationalizing the familywise error rate in terms of multiple hypotheses results in two problems. The first problem is that it leads to confusion over which hypotheses to include in the family of hypotheses that is used to compute an adjusted alpha level (Feise, 2002; O'Keefe, 2003, 2007; Trafimow & Earp, 2017). For example, a family could include all of the hypotheses in a multiway analysis of variance (ANOVA; Cramer et al., 2016), all of the hypotheses in a single study or multistudy article (sometimes called the *experimentwise error rate*), all of the hypotheses in a collection of articles that address the same issue, or even all of the hypotheses that have been and/or will be conducted by a specific researcher during their career (Hurlbert & Lombardi, 2012; O'Keefe, 2003, 2007; Trafimow & Earp, 2017). The argument that *p* values lose their meaning in exploratory analyses is based on the assumption that all of the hypotheses in a study or multistudy article should be included in a family. However, the decision to limit the definition of a family of hypotheses to those in a study or article is an arbitrary one (O'Keefe, 2003, 2007), and other arbitrary decisions lead to different conclusions regarding the meaningfulness of *p* values. For example, if the familywise error rate is based on hypotheses that are tested in a specific analysis (e.g., a multiway ANOVA; Cramer et al., 2016), then the number of hypotheses that are tested in that analysis is limited and knowable, and alpha level adjustment becomes possible, even in exploratory analyses. Alternatively, if the family of hypotheses is broadened to include all of the different hypotheses that are undertaken by a specific researcher (O'Keefe, 2003, 2007; Trafimow & Earp, 2017), then even *p* values in preregistered confirmatory analyses lose their meaning due to an unknown inflation of the alpha level based on future confirmatory tests that are undertaken by that researcher (Frane, 2015). In summary, the claim that *p* values lose their meaning in exploratory analyses and not in confirmatory analyses is only valid when the criterion for categorizing hypotheses as part of a family falls into a Goldilocks zone that is not so narrow as to make those hypotheses identifiable in exploratory analyses and not so broad as to make them unknowable in confirmatory analyses.

A second problem with the multiple hypotheses approach to the familywise error rate is that it implies that multiple null hypotheses combine to form a *universal null hypothesis*, sometimes referred to as a *composite*, *omnibus*, or *joint* null hypothesis (e.g., De Groot, 1956/2014). This universal null hypothesis predicts an overall null effect for a collection of different hypotheses (e.g., all of the hypotheses that are tested in an experiment). However, many researchers have questioned the usefulness of universal null hypotheses (Armstrong, 2014; Bender & Lange, 2001; Hurlbert & Lombardi, 2012; O'Keefe, 2003; Matsunaga, 2007; Morgan, 2007; Parascandola, 2010; Perneger, 1998; Rothman, 1990; Shulz & Grimes, 2005). One key problem is that universal null hypotheses are unlikely to be associated with theoretically-meaningful alternative hypotheses. For example, consider the four individual null hypotheses that there are no gender differences, age differences, ethnicity differences, or social class differences in self-esteem. Each of these null hypotheses may be related to theoretically-informative alternative hypotheses. However, the universal null hypothesis that there are no gender, age, ethnicity, *and* social class differences in self-esteem is difficult to interpret in the absence of any coherent theoretical framework that explains why all four of these demographic variables might be expected to exert a simultaneous influence on self-esteem.

Familywise Error Rates Based on Different Tests of the Same Hypothesis

Noting the limitations that arise from the multiple hypotheses approach, Matsunaga (2007) suggested that the familywise error rate should relate to *single* hypotheses rather than *multiple* hypotheses. Under this conceptualization, multiple testing only results in alpha inflation when it relates to the same single null hypothesis. As Matsunaga explained,

> if *multiple* H_0*s* [null hypotheses] are tested, inflation is of no concern because Type I errors are partitioned per H_0, each of which entails distinct alphas. If multiple tests are carried out *within one* H_0, however, overall Type I error rate for that H_0 becomes inflated and adjustment needs to be made (Matsunaga, 2007, p. 255).

Matsunaga's (2007) reasoning can be clarified by distinguishing between *different tests of the same hypothesis* and *single tests of different hypotheses*. To illustrate different tests of the same hypothesis, consider a researcher who tests the hypothesis that men are more prejudiced than women using five different measures of prejudice. In this case, the null hypothesis is that there is no gender difference in prejudice, and the actual alpha level for this null hypothesis is inflated as a result of the researcher conducting five different tests of this hypothesis using the five different measures of prejudice.[2] These five different tests provide the researcher with a greater number of opportunities to incorrectly reject the null hypothesis than would a single test. In contrast, if the researcher undertook a single test of the hypothesis that men are more prejudiced than women and a single test of the separate hypothesis that men have higher self-esteem than women, then the actual alpha level for each null hypothesis would be unchanged because each hypothesis is only tested once at its own prespecified nominal alpha level (Matsunaga, 2007). In other words, the researcher only has *one* opportunity to incorrectly reject the null hypothesis that men are just as prejudiced as women ($\alpha = .05$) and *one* opportunity to incorrectly reject the null hypothesis that men have the same self-esteem as women ($\alpha = .05$).

Note that Matsunaga's (2007) operationalization of the familywise error rate needs to be qualified slightly in order to be logically consistent with its premises. The qualification relates to the way in which researchers interpret the results of different tests of the same null hypothesis. If researchers require *all* of the different tests to yield a significant result in order to reject the null hypothesis, then it is not necessary to lower the alpha level of each test because this all or nothing approach means that the researcher only has *one* opportunity to reject the null hypothesis. An alpha level adjustment is only required if researchers intend to reject the null hypothesis on the basis of a significant result among *any* of the different tests that they conduct. In this case, the nominal alpha level for the null hypothesis needs to be reduced in order to take into account the fact that the researcher has increased the number of opportunities to reject the null hypothesis. It is for this reason that the familywise error rate is defined in terms of the probability of making *at least one* Type I error in a set of tests: If *any* significant result from a set of tests can be used to reject the null hypothesis, then it is important to know the probability of the occurrence of *any* Type I error in that set of tests.

Critics of Matsunaga's (2007) approach might argue that researchers who conduct single tests of multiple hypotheses have a greater probability of incorrectly rejecting one of the null hypotheses that they test (i.e., making a Type I error). This is correct, but only insofar as those researchers also have a greater probability of incorrectly accepting, correctly accepting, and correctly rejecting one of their null hypotheses. These increased probabilities are due to testing more hypotheses rather than to any changes in the probabilities of correctly or incorrectly accepting or rejecting any individual hypothesis. To illustrate, consider a gambler who purchases 100 lottery tickets. Although this mass purchase increases the probability that the gambler will win the lottery, it does not increase or decrease the probability that any one of her lottery tickets will be the winning ticket. In the same way, a researcher who undertakes single tests of 100 different null hypotheses will have a relatively high probability of incorrectly rejecting *one* of those hypotheses, but she will not increase the probability of incorrectly rejecting *each* hypothesis.

O'Keefe (2007) criticized Matsunaga's (2007) approach by arguing that any test of a null hypothesis may be reconstrued as one of a collection of tests of that same null hypothesis that have been conducted in past studies and/or that will be conducted in future studies. Consequently, like the multiple hypotheses approach, Matsunaga's multiple testing approach may lead to confusion about what to include in a family of tests. However, O'Keefe's criticism does not appear to be well-founded. There are two situations in which past and/or future studies may test the same null hypothesis, and neither situation results in an inflation of the alpha level. In the first situation, the same test of a null hypothesis is repeated in an exact replication. In this case, the alpha level is not inflated because the Type I error rate already reflects the probability of incorrectly rejecting a null hypothesis *in a long run of exact replications.* Consequently, the Type I error rate remains constant if researchers simply repeat the same test over and over again using different samples that have been randomly drawn from the exact same population. However, this first situation is somewhat hypothetical and may even be regarded as impossible in the social sciences because populations of people change over time and location (e.g., Gergen, 1973; Iso-Ahola, 2017; Schneider, 2015; Serlin, 1987; Stroebe & Strack, 2014). Yesterday's population of psychology undergraduate students from the University of Newcastle, Australia will be a different population to today's population of psychology undergraduate students from the University of Newcastle, Australia. In particular, today's population will have aged, been exposed to new information (e.g., a new lecture, news of a terrorist attack), and/or experienced new events (e.g., a weather-related event, an examination). Consequently, a second situation of inexact replication needs to be considered. In this situation, each new sample represents a different population and, therefore, a different null hypothesis. As I have explained above, the alpha level is not inflated when researchers test different null hypotheses (Matsunaga, 2007). Consequently, alpha adjustment is also unnecessary in this second situation. Alpha adjustment is only necessary when different tests investigate the *same* null hypothesis within samples that are randomly drawn from the *exact same population*.

Hence, alpha adjustment is usually only necessary within the same study, and tests conducted in actual past or future studies are not relevant.

O'Keefe (2007) would further argue that a cumulative science is not possible if researchers always test different null hypotheses. However, tests of several different statistical null hypotheses can be used to make inferences about the same substantive alternative hypothesis. For example, the prediction that there is no significant difference in prejudice between male and female psychology students at the University of Newcastle, Australia draws from a different population and represents a different null hypothesis to the prediction that there is no significant difference in prejudice between male and female psychology undergraduates at the University of Newcastle-upon-Tyne, UK. Nonetheless, the results of tests of both null hypotheses can be used to make inferences about the broader substantive alternative hypothesis that men show greater prejudice than women. It is important to note here that the nominal alpha level refers to the probability of incorrectly rejecting the statistical null hypothesis, not the substantive alternative hypothesis. Consequently, no alpha adjustment is required when considering multiple inferences to substantive alternative hypotheses.

Implications for the Interpretation of *p* Values in Exploratory Analyses

Following Matsunaga (2007), I have proposed that it is more appropriate to operationalize the familywise error rate in relation to different tests of the same hypothesis in the same study rather than multiple tests of multiple hypotheses. Matsunaga's approach has two implications for interpreting *p* values in exploratory analyses. The first is that alpha level adjustment is less necessary in exploratory analyses than might otherwise be the case, and the second is that even when different tests of the same hypothesis necessitate an alpha level adjustment, the localized hypothesis-bound nature of this testing makes the basis for such adjustments objectively knowable and transparent in exploratory analyses. I discuss each of these issues in turn.

Alpha Level Adjustment is Not as Necessary in Exploratory Analyses

As indicated earlier, it is not necessary to lower the nominal alpha level when undertaking single tests of several different hypotheses. Exploratory analyses often involve many tests of this type. Consequently, alpha level adjustments are less necessary in exploratory analyses than would be the case if researchers adopted a multiple hypotheses approach to the familywise error rate.

Again, it remains the case that the more hypotheses that a researcher tests, the greater probability that they will make a Type I error. However, this increased probability is distributed across the entire collection of hypotheses that are tested rather than localized to any one specific hypothesis. Consequently, it does not threaten the validity of any single test. If (a) researchers are interested in determining whether evidence supports or falsifies specific hypotheses rather than amorphous collections of hypotheses (i.e., universal null hypotheses), and (b) the probability that one hypothesis is true does not influence the probability that another hypothesis is true, then there is no need to adjust alpha levels for single tests of multiple hypotheses.

The Extent to Which Alpha Levels Should be Adjusted is Objectively Verifiable

If the familywise error rate is restricted to different tests of the same hypothesis in the same study, then readers of exploratory research reports can objectively verify the precise number of tests that researchers have undertaken in relation to each hypothesis. To illustrate, reconsider the researcher who tests the hypothesis that men are more prejudiced than women. If the researcher's study contains five different measures of prejudice, then the corresponding null hypothesis has the potential to be associated with five different tests, and the familywise alpha should be adjusted accordingly (e.g., from .05 to .01). Note that, for this approach to be tenable, it is necessary for researchers who undertake exploratory analyses to report all of the variables that they include in their research (Simmons, Nelson, & Simonsohn, 2011). Hence, openness and transparency are essential when undertaking exploratory research.

Counterarguments Considered

It could be argued that if Matsunaga's (2007) approach to familywise error is adopted, then researchers who undertake exploratory analyses may be tempted to artificially dissect their null hypotheses on a post hoc basis in order to give the impression of several independent null hypotheses that require no alpha adjustment. However, this deceptive approach is unlikely to be tenable in practice. To illustrate, consider our prejudice researcher once again. Using an adjusted alpha level of .01, he might find that men are more prejudiced than women (*p*s < .01) on all but one of his five measures of prejudice, and that the *p* value for the nonsignificant measure is .02. He might also recognize that maintaining his alpha level at an unadjusted level of .05 rather than .01 allows him to reclassify the result for this fifth measure as being "significant." However, to justify a lack of adjustment to his familywise alpha level, he would need to provide five *specific independent* theoretical rationales to explain why men would be expected to show higher prejudice than women on each separate prejudice measure. It is unlikely that he would be able to construct a convincing theoretical argument to support the development of these five independent hypotheses.

A second potential problem with adjusting the alpha level based on the familywise error rate in exploratory analyses is that this approach is susceptible to *p*-hacking (Simmons et al., 2011). For example, a researcher may conduct several slightly different tests of the same null hypothesis and obtain a mixture of significant and nonsignificant results. The researcher may then only report the results of a single test that yielded a significant result and omit any mention of any of the other tests. This approach would allow the researcher to give the false impression that an unadjusted alpha level is appropriate.

Preregistration offers an effective solution to the *p*-hacking problem. However, alternative solutions are available for researchers who wish to undertake non-preregistered research. Specifically, researchers can (a) list all of the variables that they included in their research study (Simmons et al., 2011) and (b) undertake a sensitivity analysis (e.g., Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2014; Thabane et al., 2013). The sensitivity analysis would involve reporting the results of variations in the analyses that a potential *p*-hacker might employ (e.g., with and without outliers, with and without covariates, with and without transformations) in order to demonstrate that the key pattern of significant and nonsignificant results remains stable.

A third potential problem with the proposed approach to multiple testing relates to the issue of sample-contingent data analysis rules or *forking paths* (Gelman & Loken, 2014). Sample-contingent data analysis rules take the form of "if…then" conditions that refer to information in the current sample of data in order to determine whether or not to undertake a specific analysis. For example, a sample-contingent rule might be: "if a participant's data value is greater than three standard deviations from the sample mean, then exclude that participant from the test. Otherwise, retain the participant." Other examples include undertaking a nonlinear transformation of a variable if its distribution exceeds a certain threshold value in the sample (e.g., +/-2.0 skewness) and including covariates in tests if they are empirically associated with outcome variables in the sample. These sample-contingent data analysis rules contribute to the multiple testing problem because they result in different tests being conducted in different samples in the hypothetical long run of exact replications. For example, a null hypothesis may be tested using a transformed variable in a sample in which the variable is skewed but a nontransformed variable in a sample in which it is not skewed. In this case, *two* different tests are used to investigate the same null hypothesis in the hypothetical long run of exact replications. This "invisible multiplicity" (Gelman & Loken, 2014, p. 460) leads to alpha level inflation.[3]

If the familywise error rate is operationalized on the basis of tests of multiple hypotheses, then it becomes impossible to count the number of sample-contingent data analysis rules that are used in an exploratory study because not all tests and hypotheses are documented. However, if the familywise error rate is restricted to different tests of the same null hypothesis, then it is much easier to identify and account for the sample-contingent data analysis rules that apply to that hypothesis. For example, if an outcome variable is transformed because it is skewed in the sample, then it should be counted as being involved in two different tests when computing the familywise error rate: one in which it is transformed and one in which it is not. Note that sample-contingent data analysis rules do not need to be preregistered in order to be identifiable. In an exploratory research report, sample-contingent data analysis rules are any rules or procedures that are described in the research report that have the potential to lead to different tests in different samples.

Conclusion

In summary, several researchers have argued that *p* values lose their meaning in exploratory analyses (Bender & Lange, 2001; Dahl et al., 2008; De Groot, 1956/2014; Forstmeier et al., 2016; Nosek et al., 2017; Nosek & Lakens, 2014; Wagenmakers, 2016). However, this argument is only tenable under a particular operationalization of the familywise error rate – one that is based on multiple tests of multiple hypotheses. This operationalization is inappropriate because (a) it leads to arbitrary decisions about which hypotheses should be counted in computations of the familywise error rate, and (b) it relates to a theoretically implausible universal null hypothesis. Following Matsunaga (2007), I have argued that the "family" that is used to estimate the familywise error rate should consist of different tests of the *same* hypothesis. This conceptualization is more appropriate because (a) it demarcates alpha levels to tests that are conducted in the same study, and (b) it does not implicate a universal null hypothesis.

Critically, if the familywise error rate is based on different tests of the same hypothesis in the same study, then alpha level adjustments in exploratory analyses are (a) less necessary and (b) objectively verifiable. Based on these points, *p* values do not necessarily lose their meaning in exploratory analyses due to an unknown inflation of the alpha level, and the validity of thousands of exploratory research studies that have used *p* values is not necessarily in jeopardy for this particular reason.

To end, I should note that I have not discussed the ongoing debate about the meaning of *p* values in general (e.g., Amrhein et al., 2017; Bradley & Brand, 2016; Nickerson, 2000). However, it is worth noting that part of the replication crisis can be attributed to a tendency for researchers to misinterpret the meaning *p* values. Two such misinterpretations are particularly problematic in this respect.

First, *p* values only give an indication of the replicability of an effect insomuch as they give an indication of the reality of the effect in a particular context and sample, because only real effects have the potential to be replicated. However, a lot of real effects are limited to specific contexts and populations. Hence, the only valid method of establishing whether or not an effect is replicable in a given context and population is to attempt to replicate it in that context and population.

Second, *p* values do not indicate whether or not a substantive alternative hypothesis is true. A statistical null hypothesis may be false for many different reasons, only one of which represents the substantive alternative hypothesis. Hence, the current focus on Type I error needs to be balanced with a concern about *inferential error* in order for researchers to arrive at correct conclusions about substantive alternative hypotheses.

In summary, *p* values have a relatively limited meaning with regard to replicability and the truth of substantive alternative hypotheses. My principle argument in this article has been that what little meaning they do have can be retained in exploratory research situations.

References

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *PeerJ 5*, e3544 doi: 10.7717/peerj.3544

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*, 502-508. doi: 10.1111/opo.12131

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*, *54,* 343-349. doi: 10.1016/S0895-4356(00)00314-0

Bradley, M. T., & Brand, A. (2016). Significance testing needs a taxonomy or how the Fisher, Neyman–Pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports*, *119,* 487-504. 10.1177/0033294116662659

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*, 640-647. doi: 10.3758/s13423-015-0913-5

Dahl, F. A., Grotle, M., Benth, J. Š., & Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. *European Journal of Epidemiology*, *23,* 237-242. doi: 10.1007/s10654-008-9230-x

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research. Translated and annotated by Wagenmakers, E. J., Borsboom, D., Verhagen, J., Kievit, R., Bakker, M., Cramer, A.,…van der Maas, H. L. J. *Acta Psychologica*, *148,* 188-194. doi: 10.1016/j.actpsy.2014.02.001

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology 2*: 8. doi: 10.1186/1471-2288-2-8

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*. doi: 10.1111/brv.12315

Frane, A. V. (2015). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice*, *1*, 2.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102,* 460. doi: 10.1511/2014.111.460

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26,* 309-320. doi: 10.1037/h0034436

Hurlbert, S. H., & Lombardi, C. M. (2012). Lopsided reasoning on lopsided tests and multiple comparisons. *Australian & New Zealand Journal of Statistics*, *54*, 23-42. doi: 10.1111/j.1467-842X.2012.00652.x

Iso-Ahola, S. E. (2017). Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology*, *8:*879. doi: 10.3389/fpsyg.2017.00879

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. *Communication Methods and Measures*, *1*, 243-265. doi: 10.1080/19312450701641409

Morgan, J. F. (2007). *P* value fetishism and use of the Bonferroni adjustment. *Evidence-Based Mental Health*, *10*, 34-35. doi: 10.1136/ebmh.10.2.34

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1,* 0021. doi: 10.1038/s41562-016-0021

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301. doi: 10.1037//1082-989X.S.2.241

Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. (2017). The preregistration revolution. Retrieved from https://osf.io/54n36/

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137-141. doi: 10.1027/1864-9335/a000192

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? *Human Communication Research, 29*, 431-447. doi: 10.1111/j.1468-2958.2003.tb00846.x

O'Keefe, D. J. (2007). Responses to Matsunaga: It takes a family-a well-defined family-to underwrite familywise corrections. *Communication Methods and Measures*, *1*, 267-273. doi: 10.1080/19312450701641383

Parascandola, M. (2010). Epistemic risk: Empirical science and the fear of being wrong. *Law, Probability & Risk*, *9*, 201-214. doi: 10. 1093/lpr/mgq005

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*, 1236-1238.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1,* 43-46.

Schulz, K. F., & Grimes, D. A. (2005). Multiplicity in randomised trials I: Endpoints and treatments. *The Lancet*, *365*, 1591-1595. doi: 10.1016/S0140-6736(05)66461-6

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology*, *34*, 365-371. doi: 10.1037/0022-0167.34.4.365

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366. doi: 10.1177/0956797611417632

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*, 411-432. doi: 10.1007/s11192-014-1251-5

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702-712. doi: 10.1177/1745691616658637

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59-71. doi: 10.1177/1745691613514450

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., ... & Debono, V. B. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC medical research methodology*, *13*, 92. doi: 10.1186/1471-2288-13-92

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*, *45*, 19-27. doi: 10.1016/j.newideapsych.2017.01.002

Vasilopoulos, T., Morey, T. E., Dhatariya, K., & Rice, M. J. (2016). Limitations of significance testing in clinical research: A review of multiple comparison corrections and effect size calculations with correlated measures. *Anesthesia & Analgesia, 122*, 825-830. doi: 10.1213/ANE.0000000000001107

Wagenmakers, E. J. (September, 2016). [Comment]. Retrieved from https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous/comment-page-1#.WR15rHrME71

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632-638. doi: 10.1177/1745691612463078

Endnotes

1. In the present article, I refer to the popular and contemporary hybrid model of null hypothesis significance testing that incorporates aspects of both the Neyman-Pearson and Fisherian approaches. This hybrid model suffers from several limitations that often lead to a misinterpretation of *p* values. For recent discussions, please see Amrhein, Korner-Nievergelt, & Roth (2017) and Bradley and Brand (2016).
2. In this example, the outcome variables are likely to be correlated with one another. This lack of independence between tests of the same null hypothesis is likely to be quite common. For a recent overview of methods of controlling for the familywise error rate in such cases, please see Vasilopoulos, Morey, Dhatariya, and Rice (2016).
3. As discussed in relation to O'Keefe's (2007) criticisms, it is usually inappropriate to consider tests from *actual* past and/or future studies when computing the familywise error rate because such studies are likely to represent inexact replications that draw samples from *different populations* and so test different null hypotheses. In contrast, it is appropriate to consider sample-contingent data analysis rules when computing the familywise error rate because these rules refer to a *hypothetical* long run of replications in which it is possible to conduct different tests of the same null hypothesis in different studies that have randomly drawn their samples from the *exact same population*.

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.