An Evaluation of Four Solutions to the Forking Paths Problem: Adjusted Alpha, Preregistration, Sensitivity Analyses, and Abandoning the Neyman-Pearson Approach

Mark Rubin
The University of Newcastle, Australia

Citation: Rubin, M. (2017). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology, 21*, 321-329. doi: 10.1037/gpr0000135

Abstract

Gelman and Loken (2013, 2014) proposed that when researchers base their statistical analyses on the idiosyncratic characteristics of a specific sample (e.g., a nonlinear transformation of a variable because it is skewed), they open up alternative analysis paths in potential replications of their study that are based on different samples (i.e., no transformation of the variable because it is not skewed). These alternative analysis paths count as additional (multiple) tests and, consequently, they increase the probability of making a Type I error during hypothesis testing. The present article considers this *forking paths problem* and evaluates four potential solutions that might be used in psychology and other fields: (a) adjusting the prespecified alpha level, (b) preregistration, (c) sensitivity analyses, and (d) abandoning the Neyman-Pearson approach. It is concluded that although preregistration and sensitivity analyses are effective solutions to *p*-hacking, they are ineffective against result-neutral forking paths, such as those caused by transforming data. Conversely, although adjusting the alpha level cannot address *p*-hacking, it can be effective for result-neutral forking paths. Finally, abandoning the Neyman-Pearson approach represents a further solution to the forking paths problem.

Keywords: forking paths; null hypothesis significance testing; preregistration; replication crisis; sensitivity analyses

Copyright © 2017, American Psychological Association. This self-archived article is provided for non-commercial and scholarly purposes only.

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

Imagine that a researcher predicts that men have higher self-esteem than women. Following data collection, the researcher observes that self-esteem scores in the sample are negatively skewed and, therefore, not ideal for a parametric test. Consequently, the researcher decides to $\log_{10}$ transform the self-esteem scores in order to produce a more normal distribution. Although this transformation makes the self-esteem scores more suitable for a *t* test, it also causes a statistical problem that increases the probability of incorrectly rejecting the null hypothesis that there is no gender difference in self-esteem (i.e., making a Type I error).

This example represents part of a broader problem that Gelman and Loken (2013, 2014) have described as the *garden of forking paths*. Forking paths represent tests that are based on different analytical approaches, such as transforming or not transforming self-esteem scores. In the above example, the researcher followed only one analysis path (transform the scores) because it was considered to be the best (most valid) path to take. Nonetheless, the fact that a second potential analysis path exists (do not transform the scores) increases the probability of making a false positive (Type I) error during hypothesis testing.

In their articles on this subject, Gelman and Loken (2013, 2014) provided some useful illustrations of the forking paths problem using examples from several published research articles. However, they did not provide an in-depth discussion of potential solutions to the problem. In the present article, I explain the forking paths problem and then critically evaluate four potential solutions that might be used in psychology and other fields: (a) adjusting the alpha level, (b) preregistration, (c) sensitivity analyses, and (d) abandoning the Neyman-Pearson approach. I explain why preregistration and sensitivity analyses are ineffective solutions and why adjusting the alpha level and abandoning the Neyman-Pearson approach are more effective solutions.

What is a Forking Path?

According to Gelman and Loken (2014), a forking path occurs when researchers compute "a single test based on the data, but in an environment where a different test would have been performed given different data" (p. 460). More formally, a forking path is an "if…then" decision rule that is used to determine the type of test to be undertaken. The criteria for implementing this rule are idiosyncratic aspects of the sample. Hence, a sample-contingent analysis rule might take the form: "If certain values of data exceed threshold X, then conduct Test X. Otherwise, conduct Test Y." As a more concrete example, a forking path might take the form: "If self-esteem scores have ±2.0 skewness, then $\log_{10}$ transform them. Otherwise, leave the scores nontransformed."

Sample-contingent analysis rules result in different analyses for different samples given that, by definition, different samples vary in their idiosyncratic characteristics. For example, self-esteem scores may be skewed beyond a certain level in some samples but not in others, leading to tests that operate on transformed data in some samples and tests that operate on nontransformed data in others. This issue is exacerbated for smaller samples because smaller samples are less representative of the population and, consequently, more likely to contain idiosyncratic characteristics (Gelman & Loken, 2013).

Forking paths are not limited to nonlinear transformations. They can refer to any sample-contingent basis for determining different analyses. A nonexhaustive list of forking paths includes:

1. only excluding participants if their data exceeds a threshold criterion value relative to other participants in the sample (e.g., ±3.0 *SD*s from the sample mean);
2. only excluding variables (e.g., scale items) if they exceed a threshold criterion value relative to other variables in the sample (e.g., exploratory factor analyses and/or internal reliability analyses);
3. only undertaking a nonlinear transformation if the data fails to conform to a certain threshold distribution in the sample (e.g., ±2.0 skewness);
4. only conducting a particular statistical test if data meet a sample-specific criterion (e.g., heterogeneity of variance);
5. only including variables as covariates in an analysis if they are empirically associated with predictor or outcome variables in the sample;
6. only testing an interaction effect if a relevant main effect is nonsignificant in the sample; and
7. only testing simple main effects if a relevant interaction effect is significant in the sample.

Figure 1 provides an example of how three of these forking paths can result in eight potential tests. The green boxes in Figure 1 represent the analysis that a researcher actually conducted in their study. The blue boxes represent other potential analysis paths that the researcher did not follow but that could be followed in any subsequent replications of the same study.

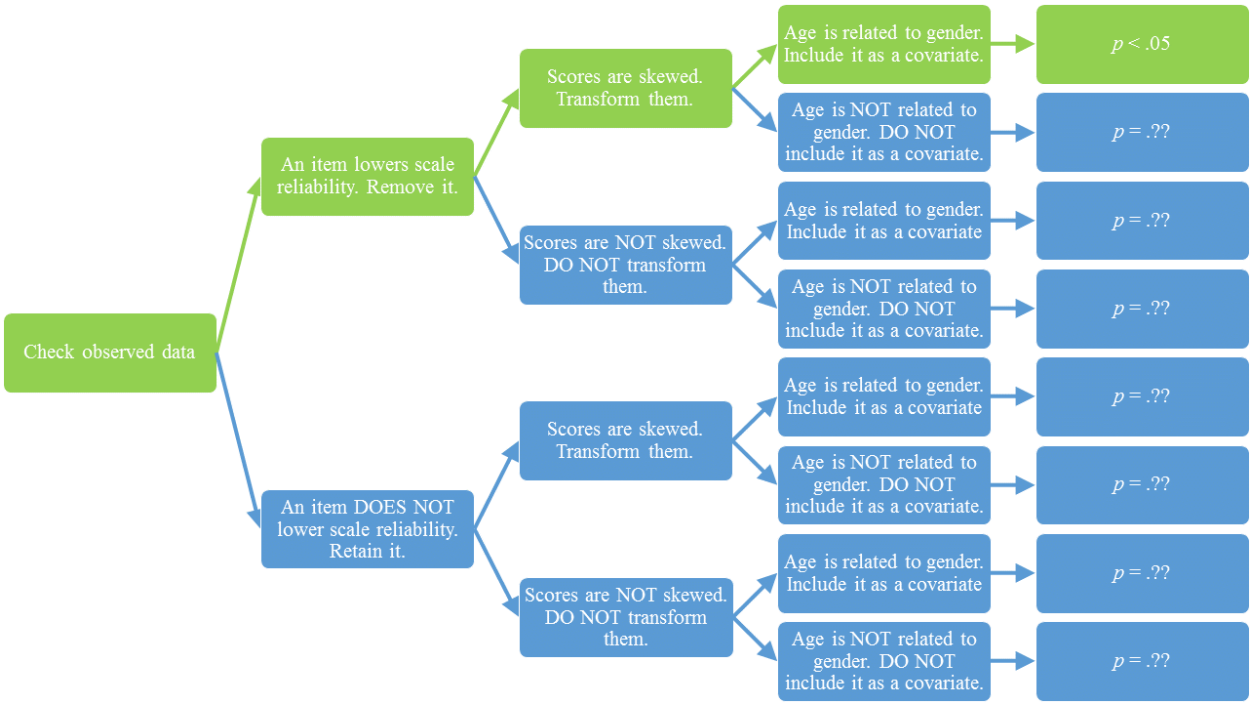

Figure 1. An illustration of three forking paths in which decisions about removing items, transforming scores, and including covariates are based on the idiosyncrasies of a specific sample of data.

Why are Forking Paths Problematic?

Forking paths are problematic because they increase the number of potential tests in a study's analysis protocol and, consequently, they increase the probability of making a Type I error. To explain, it is necessary to take a step back and consider the Neyman and Pearson (1933) approach to null hypothesis significance testing.[1] According to this approach, in order to determine if a *p* value is significant or nonsignificant, its value needs to be judged against a prespecified threshold value (e.g., .05). This *alpha level* refers to the probability of making a Type I error in a long run of exact replications of that study, with each replication randomly sampling data from the same population. Following this interpretation, if a *p* value falls below the prespecified alpha level (i.e., $p < .05$) then, assuming that the null hypothesis is true, a maximum of 5% of samples in the long run of exact replications will yield a result as extreme or more extreme as that observed in the original study. Importantly, the decision about whether or not to classify a result as being "significant" does not apply to the single result from the original study but rather to the series of results from the long run of exact replications of that study. These exact replications do not need to be actually conducted in order to decide whether or not a result is significant. However, they do need to be taken into account when specifying the probability of making a Type I error regarding this decision.

Forking paths are problematic for the Neyman-Pearson approach because they imply that different tests will be conducted in different exact replications of the same study. For example, if a researcher transforms self-esteem scores because they are skewed in their original study, then they are making an implicit commitment to follow the same conditional rule in subsequent replications based on different samples (i.e., transform self-esteem scores when they are skewed but not when they are not skewed). In the context of each specific sample, this is a reasonable and justifiable rule to follow. Nonetheless, in the context of a long run of exact replications based on different samples, this conditional rule results in two

different tests: one based on transformed scores and the other based on nontransformed scores. Hence, this conditional rule represents a forking path in the analysis protocol because it leads to different tests in different replications of the same study.

The occurrence of different tests in different replications increases the probability of making a Type I error due to multiple testing (de Groot, 2014; Frane, 2015; Szucs, 2016). Note that this multiple testing problem occurs even if a researcher has conducted only a single test in their original study. If the choice of that test is determined by the idiosyncrasies of the sample, then the test may change from one sample to the next in the long run, and it will count as two potential tests in the analysis protocol for exact replications of that study.[2] For example, in Figure 1, the researcher followed only one analysis path. Nevertheless, the three sample-contingent analysis decisions in their analysis protocol result in eight potential tests in replications of the study, and these multiple tests increase the probability of making a Type I error from 5.00% to 33.66% in a long run of exact replications of that study.

In summary, the forking paths problem is a special case of the more general multiple testing problem. However, a defining feature of the forking paths problem is that multiple tests do not need to be performed in the original study in order to be problematic. Instead, their potential presence in subsequent replications of the original study leads to an "invisible multiplicity" (Gelman & Loken, 2014, p. 460) that nonetheless increases the probability of making a Type I error. In Gelman and Loken's words,

> because the justification for *p* values lies in what would have happened across multiple data sets, it is relevant to consider whether any choices in analysis and interpretation are data dependent and would have been different given other possible data. If so, even in settings where a single analysis has been carried out on the given data, the issue of multiple comparisons emerges because different choices about combining variables, inclusion and exclusion of cases, transformations of variables, tests for interactions in the absence of main effects, and many other steps in the analysis could well have occurred with different data (p. 460).

Solutions to the Forking Paths Problem

The forking paths problem represents a common and serious threat to the validity of the Neyman-Pearson approach to null hypothesis significance testing because it increases the chances of making a Type I error. Indeed, the forking paths problem may be a contributor to the current replication crisis in psychology (Gelman & Loken, 2013, 2014; Open Science Collaboration, 2015). Consequently, it is important to consider methods for diminishing or avoiding this problem. Below, I consider four potential solutions to the forking paths problem: (a) adjusting the prespecified alpha level, (b) preregistration, (c) sensitivity analyses, and (d) abandoning the Neyman-Pearson approach.

Adjusting the Alpha Level

A well-recognised solution to the general problem of multiple testing is to adjust the prespecified alpha level in order to account for the increased probability of making a Type I error (e.g., Frane, 2015; Matsunaga, 2007; Rubin, 2017). For example, if a researcher undertakes two tests of the same hypothesis, then they can use a Bonferroni adjustment to reduce the prespecified alpha level from .050 to .025 in order to maintain the actual alpha level for their hypothesis at .050. Hence, in theory, it is possible for researchers to adjust their prespecified alpha level in order to compensate for the increased Type I error probability that results from potential multiple testing in the long run based on forking paths.

One problem with the alpha adjustment solution is that it increases the Type II error probability – the probability of incorrectly accepting the null hypothesis (Matsunaga, 2007; Weber, 2007). However, there are several different approaches to adjusting the alpha level (e.g., Matsunaga, 2007; Shaffer, 1995), and they vary in the degree to which they (a) compensate for the increased Type I error probability and (b) increase the Type II error probability. Hence, researchers can choose approaches that do not have a large impact on the Type II error probability if this is a concern.

A potentially bigger problem for the alpha adjustment solution is that, in non-preregistered research studies, it is not possible to exhaustively identify all of the tests that *could* have been undertaken. The

forking paths problem implies that some of these potential tests may be conducted in some samples in the long run of exact replications. Consequently, the forking paths problem leads to an indeterminate number of potential tests in the long run, and this makes it impossible to calculate the extent to which the prespecified alpha level needs to be adjusted (e.g., de Groot, 2014; Forstmeier, Wagenmakers, & Parker, 2016; Gelman & Loken 2013, 2014).

However, *indeterminate potential testing* is only problematic for certain types of forking path. Below, I consider *result-neutral* forking paths and two types of *result-biased* forking paths – *fishing* and *p-hacking*. I show how the alpha adjustment solution is tenable for result-neutral forking paths, potentially unnecessary for fishing, and untenable for *p*-hacking.

Result-neutral forking paths. Forking paths can be based on either result-biased analysis rules or result-neutral analysis rules. Result-biased analysis rules specify an iterative form of testing that continues until a significant result is obtained. Consequently, result-biased forking paths produce an indeterminate number of potential tests in the long run because it is unclear how many tests will need to be conducted in subsequent samples in order to obtain a significant result.

In contrast, result-neutral analysis rules specify a noniterative form of testing that is based on logical and/or conventional analytical principles. Consequently, result-neutral forking paths generate a fixed and knowable number of potential tests in the long run. For example, the analysis rule "if self-esteem scores have ±2.0 skewness, then $\log_{10}$ transform them" generates only two potential tests in the long run: one in which scores are transformed and one in which they are not transformed.

The alpha adjustment solution cannot be applied to result-biased forking paths because they produce an indeterminate number of potential tests in the long run. However, it can be applied to result-neutral forking paths because they produce a fixed and knowable number of potential tests in the long run. For example, the prespecified alpha level can be lowered to compensate for the two tests that are implied in the long run when a researcher $\log_{10}$ transforms their self-esteem data (Rubin, 2017).

It may be objected that even result-neutral analysis rules can produce an indeterminate number of potential tests in the long run because there are many potential variations to the specifications of each rule, and each variation produces more potential tests. For example, the threshold value for the analysis rule "exclude self-esteem data that exceeds ±3.0 *SD*s from the sample mean" has the potential to be varied from ±3.0 *SD*s to ±2.5 *SD*s, ±2.0 *SD*s, or any other value. However, these potential variations do not affect the Type I error probability for the original ±3.0 *SD*s test because they do not represent exact replications of that test. If a researcher actually used a threshold value of ±3.0 *SD*s in their original test, then a potential variation of that test that used a threshold value of ±2.5 *SD*s would represent an inexact replication of the original test, and so it would not count as part of the long run of exact replications of that test. Consequently, potential variations of actual analysis rules do not inflate the Type I error probability associated with those rules. If this was not the case, then even preregistered analysis rules would be susceptible to Type I error inflation because a large number of potential variations to those rules remains possible outside of the preregistered testing protocol (e.g., Silberzahn et al., 2017).

It might also be objected that researchers may not report some actual analysis rules. For example, researchers might report that they excluded outliers using the ±3.0 *SD*s rule but fail to report that they also tried excluding outliers using a ±2.5 *SD*s rule in order to see how this alternative test affected their results. However, this approach would represent a result-biased analysis rule (*p*-hacking) rather than a result-neutral analysis rule. I discuss result-biased rules below.

In summary, the alpha adjustment solution is tenable for result-neutral forking paths, such as nonlinear transformations or outlier exclusions, because they produce a fixed and knowable number of potential tests in the long run. In contrast, result-biased forking paths produce an indeterminate number of potential tests in the long run and, consequently, the alpha adjustment solution is not possible for this type of forking path. However, although alpha adjustment is not possible for result-biased forking paths, it may not always be necessary. In particular, alpha adjustment can be unnecessary in the case of fishing.

Fishing. Fishing involves testing many different hypotheses in an effort to find a significant result. This result-biased analysis strategy produces an indeterminate number of potential tests in the long run of exact replications because it is unclear how many different hypotheses the researcher would need to test in future samples in order to obtain a significant result. Hence, it is not possible to compute an alpha adjustment to compensate for fishing (e.g., de Groot, 2014; Forstmeier, Wagenmakers, & Parker, 2016; Gelman & Loken 2013, 2014). But does fishing always require an alpha adjustment? The answer depends on how researchers operationalize the *familywise error rate*.

The familywise error rate is the probability of making at least one Type I error in a series (or "family") of statistical tests. There are two ways of operationalizing the familywise error rate. The first way is in terms of multiple tests of *several different* hypotheses (e.g., all of the hypotheses in a study; de Groot, 2014). In this case, the familywise error rate refers to the probability of making a Type I error when testing a *joint null hypothesis* that any of the hypothesised effects is zero. The theoretical implications of rejecting this joint null hypothesis are unclear when its constituent individual hypotheses are derived from different theories (Rubin, 2017). It is more meaningful to test a joint hypothesis when its constituent hypotheses are derived from the same theory because, in this case, its rejection can be taken as providing support for the theory (Hewes, 2003). However, even under these circumstances, it is only necessary to adjust the alpha level for each constituent hypothesis if a significant result (i.e., $p < \alpha_{adjusted}$) for *any one* of the constituent hypotheses is regarded as being sufficient to reject the joint null hypothesis as a whole (Matsunaga, 2007; Rubin, 2017; Weber, 2007).[3] Although this approach may be appropriate in some fields, psychology researchers do not normally investigate theories (i.e., collections of hypotheses) in this way. Instead, they usually test theories on a hypothesis-by-hypothesis basis in order to understand which hypotheses have been confirmed and which have been disconfirmed and may need revising or rejecting (Lakatos, 1976). Hence, even when it is theoretically-meaningful to test a joint null hypothesis, most psychology researchers are likely to test each individual hypothesis separately, at its own individual unadjusted alpha level, rather than jointly at an adjusted alpha level.

The second way of operationalizing the familywise error rate is in relation to multiple tests of the *same individual* hypothesis (Matsunaga, 2007; Rubin, 2017). In this case, an alpha adjustment is only required to correct for the inflated error rate caused by multiple tests of that hypothesis; not multiple tests of a joint hypothesis that comprises several different hypotheses.

The way in which the familywise error rate is operationalized has important implications for alpha adjustment in the context of fishing. If familywise error refers to several different hypotheses, then alpha adjustment is necessary in order to compensate for the potential testing of different hypotheses that results from fishing. However, it is not possible to make this adjustment because the number of potential tests is unknown. Consequently, the alpha adjustment solution fails in this case (de Groot, 2014; Gelman & Loken, 2013, 2014). In contrast, if familywise error is limited to tests of the same individual hypothesis (Matsunaga, 2007; Rubin, 2017), then alpha adjustment is not necessary to compensate for potential tests of different hypotheses in the long run. Consequently, alpha adjustment is not necessary in the context of fishing.

To illustrate, consider a researcher who follows the fishing strategy by testing Hypothesis 1 first and then only proceeding to test Hypotheses 2 to X (some undetermined number) if they find no significant result in relation to Hypothesis 1. If familywise error refers to several different hypotheses (i.e., Hypotheses 1 to X), then this result-biased analysis strategy represents a problematic forking path in the long run of exact replications because, although Hypotheses 2 to X may not be tested in the original study, one or more of them may be tested in subsequent studies that yield nonsignificant results for Hypothesis 1. Consequently, the familywise error rate is inflated to an unknown extent, and it cannot be adjusted appropriately. In contrast, if the familywise error rate is limited to multiple tests of the same individual hypothesis, then the probability of making a Type I error for Hypothesis 1 is unrelated to the probability of making Type I errors for any of Hypotheses 2 to X (Matsunaga, 2007). In this case, each hypothesis is tested at its own individual alpha level rather than in relation to a joint null hypothesis. Consequently, although one or more of Hypotheses 2 to X may be tested in the long run, this analytical variability does not inflate the Type I error probability of the test of Hypothesis 1.

In summary, if the familywise error rate is limited to multiple tests of the same individual hypothesis, then neither the actual testing nor the potential testing of different hypotheses inflates the Type I error probabilities of each individual hypothesis. Consequently, no alpha adjustment is required is required for fishing. Alpha adjustment is only required for actual or potential multiple tests of the same individual hypothesis.[4]

***p*-hacking.** Fishing involves the actual and potential testing of several different hypotheses with the aim of finding a significant result. In contrast, *p*-hacking involves the actual and potential testing of the same individual hypothesis with the aim of finding a significant result (Simmons, Nelson, & Simonsohn, 2011). For example, a researcher might investigate the hypothesis that men have higher self-esteem than women by conducting tests with and without outliers, with and without various transformations, and with and without various covariates. They might find a significant result only when they $\log_{10}$ transform the data and exclude outliers at 2.5 *SD*s, and so they might only report that result and conceal their other nonsignificant results.

Like fishing, *p*-hacking is based on a result-biased analysis rule in which an iterative testing procedure only ceases once the desired result has been obtained. Consequently, like fishing, *p*-hacking leads to an indeterminate number of potential tests in the long run of exact replications. However, unlike fishing, *p*-hacking involves conducting multiple tests of the same individual hypothesis, rather than different hypotheses that comprise a joint hypothesis. Consequently, even when familywise error rate is operationalized in terms of the same individual hypothesis, *p*-hacking results in an indeterminate number of potential tests in the long run of exact replications and an unknown inflation of the Type I error. Consequently, the alpha adjustment solution fails in the presence of *p*-hacking.[5]

Summary. In summary, the alpha adjustment solution is tenable for result-neutral forking paths (e.g., "exclude self-esteem data if it exceeds ±3.0 *SD*s from the sample mean"). Furthermore, if familywise error is limited to the same individual hypothesis, then alpha adjustment is not necessary for result-biased forking paths that refer to different hypotheses, such as fishing. However, the alpha adjustment solution is not tenable for result-biased forking paths that refer to the same hypothesis, such as *p*-hacking.

Preregistration

Gelman and Loken (2013, 2014) suggested preregistration as a potential solution to the forking paths problem (e.g., Nosek, Ebersole, DeHaven, & Mellor, 2017; Nosek & Lakens, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). However, this potential solution suffers from a major drawback. Specifically, if any part of a preregistered protocol describes sample-contingent "if…then" rules for analyses, then it will succumb to the same forking paths problem as non-preregistered research. For example, a preregistered protocol might specify: "If outcome scores are skewed $\geq \pm 2.0$, then $\log_{10}$ transform them. Otherwise, leave them nontransformed." Despite being preregistered, this analysis rule represents a forking path that leads to multiple testing in the long run and increases the Type I error probability. Hence, preregistration per se does not mitigate against the forking paths problem.

Notably, advocates of preregistration encourage the explicit identification of sample-contingent analysis rules in analysis protocols. For example, in their discussion of a template for preregistered research in social psychology, van't Veer and Giner-Sorolla (2016) suggested that researchers specify contingencies for (a) handling outliers, (b) removing items that reduce reliability, (c) transforming variables to achieve more normal distributions, and (d) statistical corrections for heterogeneity of variance. Similarly, Chambers, Feredoes, Muthukumaraswamy, and Etchells (2014) stated that:

> pre-registration does not require every step of an analysis to be specified or "hardwired"; instead, in such cases where *the analysis decision is contingent on some aspect of the data itself* then the pre-registration only requires the decision tree to be specified (e.g. "If A is observed then we will adopt analysis A1 but if B is observed then we will adopt analysis B1") (p. 12, my emphasis).

However, merely describing forking paths in a decision tree does not do anything to reduce their impact on the Type I error probability (Frane, 2015), and it is this issue that needs to be addressed in order to solve the forking paths problem.

Most recently, Nosek et al. (2017) summarised four strategies that are intended to address sample-contingent analyses in the preregistration paradigm. The first two strategies blind researchers to the substantive research results. In the first strategy, researchers first preregister their approach to data exclusions, transformations, model assumptions, etc. They then undertake their preliminary preregistered analyses on the data without proceeding to undertake more theoretically-substantive analyses. This preliminary data analysis stage allows researchers to select the most appropriate testing approach for their data. The researchers then specify this testing approach in a second preregistered analysis protocol that also includes more theoretically-substantive analyses. This two-stage sequential preregistration approach allows researchers to make decisions about data exclusions, transformations, model assumptions, etc. on the basis of their observed data without being influenced by their more substantive research results.

The second strategy is to blind researchers to the identity of different variables (e.g., hide variable names) but to allow researchers to observe the distributional characteristics of those variables (e.g., outliers, skewness). Again, this approach allows researchers to make decisions about outliers, transformations, modelling assumptions, etc. on the basis of their data without allowing them to know how those decisions impact on their more substantive research results.

Both of the above blinding strategies preclude result-biased forking paths such as *p*-hacking. However, neither strategy mitigates against result-neutral forking paths (e.g., exclude self-esteem data if it exceeds ±3.0 *SD*s from the sample mean). In other words, although both strategies prevent analyses from being influenced by the research results, neither strategy prevents analyses from being influenced by the idiosyncrasies of the data. Concealing substantive research results from researchers when they make data-contingent analysis decisions about how to handle outliers, non-normal distributions, and modelling assumptions does not prevent these decisions from creating forking paths in the long run of exact replications. Hence, although these blinding approaches are effective against *p*-hacking, they are ineffective against other, result-neutral forking paths.

Nosek et al.'s (2017) third and fourth strategies consist of documenting all of the potential forking paths that may result in a sample-contingent analysis, either by preregistering a decision tree for that analysis (see also Chambers et al., 2014) or by referring to a more general set of standard operating procedures. Again, however, merely identifying and recording forking paths in either preregistered or postregistered documents does not reduce their impact on Type I error probabilities (Frane, 2015). Instead, researchers need to adjust their prespecified alpha level in order to reduce the probability of making a Type I error in the long run.

In order to eliminate forking paths and, consequently, the need to adjust the alpha level, a preregistered analysis protocol must contain a precise description of an analysis that is to be followed regardless of the idiosyncratic characteristics of the observed sample of data. The difficulty with this approach is that any inferential benefits that are gained by eliminating forking paths must be offset against the inferential costs that are associated with an often poor quality analytical approach. In particular, if preregistered analysis protocols prevent researchers from adapting their analyses to the specific characteristics of their data, then researchers will be unable to (a) clean the data (e.g., Osborne, 2012), (b) improve the reliability and validity of their measurements (e.g., Schmidt & Hunter, 1999), (c) adjust their statistical tests in order to meet test assumptions (e.g., Erceg-Hurn & Mirosevich, 2008), and (d) control for relevant covariates (e.g., Bernerth & Aguinis, 2016). Hence, an effective preregistration solution is caught between two equally problematic threats to statistical inference: Allowing sample-contingent analysis rules in preregistered research protocols creates forking paths, but disallowing them reduces the quality of the data analyses.

In summary, although preregistration represents a tenable solution to result-biased forking paths, such as *p*-hacking, it fails to mitigate against the impact of result-neutral forking paths, such as excluding outliers.

Sensitivity Analyses

A third solution to the forking paths issue is to undertake sensitivity analyses (e.g., Thabane et al., 2013; Wasserstein & Lazar, 2016; Wigboldus & Dotsch, 2016). This approach requires researchers to be cognizant of the various forking paths that they take during their analyses and, at each fork, to investigate potential discrepancies in their results if they take the alternative path. So, for example, if the data indicate that it is appropriate to transform scores, then researchers should report the results of tests that are conducted when scores are (a) transformed and (b) not transformed. This approach allows researchers and their audience to understand the variability in their results across different analytical approaches.

Like preregistration, sensitivity analyses provide an effective solution to the *p*-hacking problem. In particular, they can be used to reveal disconfirming results that would otherwise remain undisclosed. However, sensitivity analyses are not an effective solution to the result-neutral forking paths for three reasons.

First, like preregistration, sensitivity analyses per se do not lower the prespecified alpha level. Consequently, any result that is classed as "significant" in a sensitivity analysis may be reclassified as "nonsignificant" after an appropriate alpha adjustment has been implemented. For example, a sensitivity analysis of a test with $\alpha = .05$ may return significant results of $p = .003$ when scores are transformed and $p = .030$ when scores are not transformed. Based on these consistently-significant results, the researcher might conclude that their results are robust to variations in their analytical approach. However, the second of these results would be reclassed as being nonsignificant if the alpha level was adjusted to .025 in order to account for multiple testing. Hence, consistency across results cannot be established using sensitivity analyses per se.

Second, exploring an alternative analytical path in a sample whose characteristics do not warrant that analytical path will result in a suboptimal test, and any results from that test will need to be interpreted with caution. Returning to the previous example, if data are skewed, then only the test that is conducted on the transformed data is valid, and the results of the test that is conducted on the nontransformed data must be interpreted with caution.

Finally, but most importantly, sensitivity analyses only reveal the influence of different analytical approaches in the original study and not in the long run of exact replications of that study. Consequently, they do not provide any indication of the potential variability in results in the long run.

In summary, although sensitivity analyses can be used to identify and deter *p*-hacking, they do not provide an effective solution to result-neutral forking paths because (a) they do not compensate the alpha level for the multiple testing that is associated with forking paths, (b) they include suboptimal tests, and (c) they only reveal variability in results in an original study and not in the long run of exact replications of that study.

Abandon the Neyman-Pearson Approach

The final solution to the forking paths problem is to abandon the Neyman-Pearson approach to null hypothesis significance testing. Abandoning the Neyman-Pearson approach solves the forking paths problem because forking paths are only meaningful in the context of a long run of exact replications and, in the absence of the Neyman-Pearson approach, researchers do not need to interpret their results in this context. All that matters are the actual tests that are conducted on the original sample of data, not the potential tests that might be conducted on subsequent samples of data. But, if the Neyman-Pearson approach is abandoned, then what is left? I briefly consider two potential alternatives.

The first alternative is to continue to use *p* values but to interpret them in the manner advocated by Fisher (1937, 1955, 1956) rather than Neyman and Pearson (1933). The Fisherian and neo-Fisherian approaches interpret *p* values in relation to the current sample of data rather than a long run of exact replications (Fisher, 1955; Haig, 2016; Hubbard & Bayarri, 2003; Hurlbert & Lombardi, 2009; Schneider, 2015). Consequently, Fisherian *p* values are more suitable for context-dependent studies that have less potential for exact replication, such as those conducted in psychology and the social sciences. Critically, from a Fisherian perspective, the multiple testing issue only needs to be considered in relation to tests that

are actually performed on the current data rather than tests that have the potential to be performed in a series of exact replications. Hence, forking paths are not problematic in the Fisherian paradigm.

Unlike the Neyman-Pearson approach, Fisherian *p* values are interpreted on a sliding scale of probability in which the smaller the *p* value, the less likely it is that the observed data would occur if the null hypothesis was true (Amrhein, Korner-Nievergelt, & Roth, 2017; Biau, Jolles, & Porcher, 2010; Bradley & Brand, 2016; Falissard, 2012; Haig, 2016; Hubbard & Bayarri, 2003; Wasserstein & Lazar, 2016). Hence, from a Fisherian perspective, $p = .0004$ indicates stronger evidence against the null hypothesis than $p = .04$ (Hubbard & Lindsay, 2008).

Despite this sliding scale interpretation, Fisher (1926) suggested that $p < .05$ may sometimes be a useful threshold for determining the significance of a result. However, he stressed that researchers need to make context-based interpretations of specific *p* values. As he pointed out, "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher, 1956, p. 42). This context-based approach provides a more nuanced and sophisticated approach to significance testing, in which researchers interpret the meaning of *p* values based on a range of background factors including (a) the size of the *p* value, (b) the number of tests that have been used to assess the null hypothesis, (c) the theoretical plausibility of the explanation, (d) prior evidence for the effect, (e) the effect size, (f) the variability of the effect, (g) the quality of the research methodology (e.g., sample size, measurement precision, construct validity, etc.), (h) the results of sensitivity analyses, and (i) the results of tests of alternative explanations for the effect (e.g., Amrhein et al., 2017; Wasserstein & Lazar 2016). For example, researchers should interpret a *p* value of .025 differently when it is used to infer the existence of extrasensory perception compared to when it is used to infer the existence of ethnic prejudice.

The second alternative to the Neyman-Pearson approach is more radical in that it involves abandoning *p* values altogether. The Bayesian approach replaces *p* values with Bayes factors that integrate information about the prior probability of a hypothesis based on prior theory and evidence with information about the posterior probability of the hypothesis based on the sampled data. Like the Fisherian approach, the Bayesian approach operates on a sliding scale of probability. For example, Jeffreys (1961) suggested the following scale for interpreting Bayes factors: "barely worth mentioning," "substantial," "strong," "very strong," and "decisive."

As with the Fisherian approach, the Bayesian approach only relates to the current sample of data; it is not concerned with a long run of exact replications. Consequently, it is not susceptible to the forking paths problem (Dienes, 2011; see also Bender & Lange, 2001). As Wagenmakers et al. (2014, in de Groot, 2014) explained, "what is relevant for Bayesian reasoning is not the number of tests that were executed or planned, but rather the prior belief in a particular hypothesis [Scott & Berger, 2010; Stephens & Balding, 2009]" (p. 194).

Conclusions

The forking paths problem represents a serious threat to the Neyman-Pearson approach of null hypothesis significance testing. Specifically, it can increase the probability of making a Type I error across the long run of exact replications. The present article explored four potential solutions to this problem.

Preregistration and sensitivity analyses provide effective solutions to result-biased forking paths such as *p*-hacking. However, they do not mitigate against result-neutral forking paths. If a preregistered analysis protocol includes a result-neutral analysis rule such as "exclude data if it exceeds ±3 *SD*s from the sample mean," then it will be just as susceptible to the forking paths problem as non-preregistered research. Similarly, although sensitivity analyses may be used to uncover or deter *p*-hacking, this approach does not reduce the inflated Type I error probability that results from multiple testing in the long run.

The two best solutions to forking paths relate to the prespecified alpha level. The first solution is to adjust the alpha level in order to compensate for the inflated risk of making a Type I error. This solution can be used for result-neutral forking paths (e.g., "exclude data if it exceeds ±3 *SD*s from the sample mean") because they produce a fixed and knowable number of potential tests in the long run. Alpha adjustment is not possible for fishing when familywise error refers to different hypotheses. However, it is not necessary

when familywise error refers to the same hypothesis. Finally, alpha adjustment is not possible in the context of *p*-hacking under any circumstances.

The second solution is to abandon the Neyman-Pearson approach and, consequently, to abandon concerns about the alpha level of a long run of exact replications. Two suitable alternative statistical approaches are the Fisherian and Bayesian approaches.

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

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

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research®, 468*, 885-892. doi: 10.1007/s11999-009-1164-4

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

Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, *69*, 229-283. doi: 10.1111/peps.12103

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. doi: 10.1177/0033294116662659

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*, 4-17. doi: 10.3934/Neuroscience2014.1.4

De Groot, A. D. (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

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, *6*, 274-290. doi: 10.1177/1745691611406920

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*, 591-601. doi: 10.1037/0003-066X.63.7.591

Falissard, B. (2012). Statistics in brief: When to use and when not to use a threshold p value. *Clinical Orthopaedics and Related Research*, *470*, 315-316. doi: 10.1007/s11999-011-2090-9

Fisher, R. A. (1926). *Statistical methods for research workers.* Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1937). *The design of experiments.* Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69-78.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver and Boyd.

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. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

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

Haig, B. D. (2016). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*, 489-506. doi: 10.1177/0013164416667981

Hewes, D. E. (2003). Methods as tools. *Human Communication Research, 29,* 448-454. doi: 10.1111/j.1468-2958.2003.tb00847.x

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician, 57,* 171-178. doi: 10.1198/0003130031856

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*, 69-88. doi: 10.1177/0959354307086923

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46,* 311-349.

Jeffreys, H. (1961). *The theory of probability (3rd ed.).* Oxford, UK: Oxford University Press.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed), *Can Theories be Refuted?* (pp. 205-259). Springer: Netherlands.

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

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29,* 492-510. doi: 10.1017/S030500410001152X

Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. (2017). The preregistration revolution. Retrieved from https://osf.io/2dxu5/

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. doi: 10.1126/science.aac4716

Osborne, J. W. (2012). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. London: Sage Publications.

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

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183-198.

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

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46,* 561-584. doi: 10.1146/annurev.ps.46.020195.003021

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E. C., … Nosek, B. A. (2017). *Many analysts, one dataset: Making transparent how variations in analytical choices affect results*. Retrieved from https://psyarxiv.com/qkwst/

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

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing *N*. *Frontiers in Psychology, 7*: 1444. doi: 10.3389/fpsyg.2016.01444

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

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology: A discussion and suggested template. *Journal of Experimental Social Psychology, 67,* 2-12. doi: 10.1016/j.jesp.2016.03.004

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

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician, 70,* 129-133. doi: 10.1080/00031305.2016.1154108

Weber, R. (2007). Responses to Matsunaga: To adjust or not to adjust alpha in multiple testing: That is the question. Guidelines for alpha adjustment as response to O'Keefe's and Matsunaga's critiques. *Communication Methods and Measures, 1,* 281-289. doi: 10.1080/19312450701641391

Wigboldus, D. H., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika, 81,* 27-32. doi: 10.1007/s11336-015-9445-1

Endnotes

1. The typical method of conducting null hypothesis significance tests often represents a hybrid of the Neyman-Pearson and Fisherian approaches (Amrhein et al., 2017; Biau et al., 2010; Bradley & Brand, 2016; Hubbard & Bayarri, 2003; Schneider, 2015). I discuss the Fisherian approach later on in this article. Hence, to avoid any confusion, I refer to the typical method of null hypothesis significance testing as the Neyman-Pearson approach.
2. Given that larger samples are more representative of the populations from which they are drawn, the characteristics of larger samples are more likely to reflect the characteristics of their population. Hence, if self-esteem scores are skewed beyond some threshold criterion in the first sample that is drawn from a population, then they are likely to be skewed beyond that criterion in most other large samples from that population. Consequently, if researchers use large samples, then their analytical path in exact replications is likely to be the same as that taken in the original study (e.g., self-esteem scores are skewed, and so they are transformed). However, even when using large samples, it remains possible that, in at least one case during a long run of exact replications, the relevant characteristic in the sample will be decisively different from that in the population and therefore require an alternative analysis. Based on the Neyman-Pearson approach, this single alternative analysis will then count as an additional test in the analysis protocol, and it will increase the probability of making a Type I error. In other words, although researchers may almost always turn left down a particular forking path, the fact that a right turn is possible must be taken into consideration when interpreting decisions based on null hypothesis significance tests.
3. Based on the multiplication rule for independent events, the familywise error rate is computed using the formula: $1 - (1 - \alpha)^k$, where α is the prespecified alpha level and k is the number of tests in the family. In order to compute the probability of incorrectly rejecting a joint null hypothesis, this formula multiplies the probability of incorrectly rejecting each constituent null hypothesis by the number of hypotheses in the family. However, this computation is only necessary if researchers intend to reject the joint null hypothesis on the basis of a significant result (i.e., $p < \alpha_{adjusted}$) in relation to *any one* of the constituent null hypotheses (i.e., *logical disjunction*; Weber, 2007).
4. My definition of fishing includes the strategy of testing for interaction effects when main effects are nonsignificant. Contrary to Gelman and Loken (2013, 2014), this strategy involves the testing of different hypotheses, rather than the same hypothesis, because different theoretical rationales are required in order to explain main effects and interactions (Matsunaga, 2007). For example, the hypothesis that men have higher self-esteem than women (i.e., a main effect of gender on self-esteem) will have a different theoretical explanation to the hypothesis that this gender difference is stronger among young people than among older people (i.e., an interaction between gender and age in predicting self-esteem).
5. It is important to note that even if alpha adjustment was possible to compensate for multiple testing based on *p*-hacking, *p*-hacking would remain an invalid and unethical form of data analysis because it precludes reported falsification.

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.