# *p*-Hacking Inflates Type I Error Rates in the Error Statistical Approach but not in the Formal Inference Approach

*Mark Rubin* ⏺
*Durham University, UK*

March 3, 2026

## Abstract

*p*-hacking occurs when researchers conduct multiple significance tests (e.g., $p_1;H_{0,1}$ and $p_2;H_{0,2}$) and then selectively report tests that yield desirable (usually significant) results (e.g., $p_2 \leq 0.05;H_{0,2}$) without correcting for multiple testing (e.g., $0.05/2 = 0.025$). In the present article, I consider *p*-hacking in the context of two philosophies of significance testing — the error statistical approach and the formal inference approach. I argue that although *p*-hacking inflates Type I error rates in the error statistical approach, it does not inflate them in the formal inference approach. Specifically, in the error statistical approach, the "actual" familywise error rate (e.g., $1 - [1 - 0.05]^2 = 0.098$ for two tests) is relevant because it covers both the selectively reported and unreported tests in the "actual" test procedure (i.e., $p_1;H_{0,1}$ and $p_2;H_{0,2}$). In this approach, Type I error rate inflation occurs because the "actual" error rate (0.098) is higher than the nominal error rate (0.05). In contrast, in the formal inference approach, the "actual" familywise error rate is irrelevant because (a) the researcher does not report a statistical inference about the corresponding intersection null hypothesis (i.e., $H_{0,1} \cap H_{0,2}$), and (b) the "actual" familywise error rate does not license inferences about the reported individual hypotheses (i.e., $H_{0,2}$). Instead, in the formal inference approach, only the nominal error rate is relevant, and a comparison with the "actual" error rate is inappropriate. Implications for conceptualizing, demonstrating, and reducing *p*-hacking are discussed.

*p*-hacking occurs when researchers conduct multiple significance tests and then selectively report tests that yield desired, usually significant, results without correcting for multiple testing. Hence, *p*-hacking represents a type of undisclosed cherry-picking or fishing for specific (significant) results.

Although *p*-hacking is often described as a "questionable research practice" (John et al., 2012; e.g., Nagy et al., 2025; Reis & Friese, 2022), it is widely regarded as being statistically and ethically problematic (e.g., Miller et al., 2025; Pickett & Roche, 2018; Simmons et al., 2011; Stefan & Schönbrodt, 2023). It is also believed to be a major contributor to the replication crisis (e.g., Bishop, 2019; Nagy et al., 2025). In particular, *p*-hacking is thought to inflate Type I error rates above their nominal conventional level, resulting in a larger proportion of false positive results in the literature than would otherwise be expected (e.g., Lakens, 2019, p. 221; Nagy et al., 2025; Reis & Friese, 2022; Simmons et al., 2011; Stefan & Schönbrodt, 2023). This disproportionate number of false positives is then thought to cause unexpectedly low replication rates.

In the present article, I aim to add some nuance to this view by distinguishing between two philosophies of significance testing — the error statistical approach (Mayo, 1996, 2018) and the formal inference approach (Rubin, 2021b, 2024a, 2024b). I argue that, although *p*-hacking inflates Type I error rates in the error statistical approach, it does not inflate them in the formal inference approach.

The article proceeds as follows: Section 1 explains how *p*-hacking inflates Type I error rates in the error statistical approach. Sections 2 and 3 introduce the formal inference approach and explain why *p*-hacking does not inflate Type I error rates in this approach. Section 4 considers some key areas of agreement and disagreement between the two approaches. Section 5 considers the error statistical approach's minimal severity requirement and argues that it is not necessary to obtain valid Type I error rates in the formal inference approach. Section 6 considers the nondisclosure of null results during *p*-hacking and argues that, although this practice does not inflate Type I error rates in the formal inference approach, it may bias substantive inferences under certain conditions. Finally, Section 7 summarizes the conclusions and discusses their implications for conceptualizing, demonstrating, and reducing *p*-hacking and for understanding and addressing the replication crisis.

# 1. *p*-Hacking Inflates Type I Error Rates in the Error Statistical Approach

## Type I Error Rates in the Error Statistical Approach

A Type I error rate represents the maximum frequency with which a significance tester would incorrectly reject a null hypothesis based on random sampling error (Neyman & Pearson, 1928, pp. 176–177, 231–232). Specifically, a Type I error occurs when random sampling error leads the significance tester to decide that an observed sample was not drawn from a null population when, in fact, it was. In the Neyman-Pearson approach, a Type I error "rate" refers to this decision-making performance across a hypothetical long run of repeated random sampling from a specified null population (Neyman & Pearson, 1928, p. 177).

Without denying this Neyman-Pearson long-run "performance" rationale, error statisticians adopt a "probative" approach in which they consider the error-probing capacity of a test procedure in the case at hand (Mayo, 2018, pp. 13–14, 162; Mayo & Spanos, 2006, p. 351; Mayo & Spanos, 2011, pp. 162–163). Here, a Type I error rate is used to "license" specific inferences based on the application of a particular test. As Mayo and Spanos (2006) explained, "pre-data, the choices for the type I and II errors reflect the goal of ensuring the test is capable of

licensing given inferences severely" (p. 350; see also Mayo & Spanos, 2011, p. 167). For example, a test with a nominal Type I error rate of 0.05 ($\alpha$) is capable of licensing specific statistical inferences with a minimum "worst case" severity of 0.95 (i.e., $1 - \alpha$), assuming adequate statistical assumptions (Mayo, 1996, pp. 124, 304, 399, 407; Mayo, 2008, pp. 863–864; Mayo, 2018, p. 94; Spanos & Mayo, 2015, p. 3535).

## The Problem with *p*-Hacking

As Mayo and Spanos (2011, p. 166) explained, a key requirement in the error statistical approach is that the probability of observing a significant *p*-value should be $\leq \alpha$ when all null hypotheses are true (i.e., Pr[observing a *p*-value $\leq \alpha$] $\leq \alpha$). The problem with *p*-hacking is that this probability becomes greater than $\alpha$ when multiple undisclosed significance tests are performed in the experimental testing context. Hence, when *p*-hacking occurs, the probability of making at least one Type I error becomes greater than $\alpha$.

Consequently, a test procedure that includes *p*-hacking will have an "actual" Type I error rate that is higher than the nominal "computed" error rate of $\alpha$ (Mayo, 1996, pp. 303–304, 348; Mayo, 2008, pp. 874–875; Mayo, 2018, pp. 274–275; Mayo & Cox, 2010, pp. 267–270). It is this discrepancy between "actual" and nominal error rates that is conceptualized as Type I error rate "inflation" (Mayo, 2025, p. 1).

To illustrate, imagine that a researcher conducts significance tests on 20 hypotheses $H_{0,1}$, $H_{0,2}$, $H_{0,3}$, … $H_{20}$, yielding 20 corresponding *p*-values $p_1, p_2, p_3, … p_{20}$. They then search for results that are significant at the conventional level of 0.05. They find that only $p_{13}; H_{0,13}$ is significant, and so they selectively report that particular result.[1]

In this situation, the "actual" test procedure includes searching for significance among 20 tests and then selectively reporting significant results (Mayo, 1996, p. 304). Consequently, the "actual" Type I error rate is the familywise error rate for those 20 tests (Mayo, 2018, p. 275). The upper bound of the familywise error rate can be computed using the formula $1 - (1 - \alpha)^k$, where $k$ is the number of significance tests. Hence, the "actual" error rate is 0.642 (i.e., $1 - [1 - 0.05]^{20}$). In contrast, the nominal error rate for the reported inference about $H_{0,13}$ is 0.05. Hence, the "actual" error rate is inflated above the nominal error rate, and the nominal error rate is "invalidated" or "vitiated" because it does not match the "actual" error rate (i.e., Pr[observing a *p*-value $\leq \alpha$] $> \alpha$; Mayo, 2008, p. 876; Mayo, 2018, pp. 234, 285, 438; Mayo & Spanos, 2011, p. 190).

To address this potential problem, error statisticians recommend that we check or "audit" the experimental testing context for any "biasing selection effects," such as *p*-hacking (e.g., by cross-referencing with a preregistered analysis plan; Mayo, 2018, pp. 92, 106–107, 275, 439). The "minimal severity requirement" is that we attempt to rule out ways in which a claim may be false (Mayo, 2018, p. 5). Biasing selection effects occur when hypotheses and data are generated and selected in ways that violate, secretly alter, or prevent an assessment of this requirement (Mayo, 2018, pp. 92, 269). Hence, *p*-hacking represents a biasing selection effect because its hidden multiple testing secretly inflates the "actual" Type I error rate and prevents us from effectively ruling out random sampling error. Consequently, we are left with "bad evidence, no test (BENT)" (Mayo, 2018, p. 5).

# 2. The Formal Inference Approach

The error statistical approach can be contrasted with what I call here the formal inference approach (Rubin, 2021a, 2021b, 2024a, 2024b, 2025). In this approach, relevant Type I error rates as based on formally reported (publicly specified) statistical inferences rather than "actual" test

procedures, which may be only partially reported.[2] In particular, the contents of a formally reported inference are used to determine (a) the category of statistical inference, (b) the number of significance tests associated with the inference, and (c) the specified alpha level for each test (or the conventional alpha level if none is specified). This information is then used to compute the relevant Type I error rate for the inference (e.g., Rubin, 2024b, p. 52).

## Categories of Statistical Inferences

In the formal inference approach, a statistical inference consists of (a) a provisional decision to either reject or fail to reject a formally specified statistical null hypothesis and (b) the result of one or more significance tests that indicate conditional uncertainty about this decision (Fisher, 1956, pp. 57, 108–110; Rubin, 2024b, p. 49). Three categories of statistical inference are distinguished: *individual*, *union-intersection*, and *intersection-union* (Berger, 1982; García-Pérez, 2023; Hochberg & Tamhane, 1987; Parker & Weir, 2020; Roy, 1953; Rubin, 2021b, 2024a, 2024b).

### (1) Individual Inference

If a researcher makes a decision about an *individual* alternative hypothesis (e.g., $H_{1,1}$) and corresponding *individual* null hypothesis (e.g., $H_{0,1}$) using a nominal alpha level of $\alpha$, then the Type I error rate for that decision is $\alpha$, even if the researcher makes multiple such decisions (reported or unreported) within the same study. In this case, the individual null hypothesis can only be rejected following a significant result on *one* particular test (i.e., $p_1 \leq \alpha; H_{0,1}$). Hence, there is only a *single* opportunity to make a Type I error with regard to that particular individual null hypothesis (see also García-Pérez, 2023, p. 15; Parker & Weir, 2020, pp. 563–564; Rubin, 2017a, pp. 271–272; Rubin, 2020, p. 380; Rubin, 2021a; Rubin, 2021b, pp. 10978–10983; Rubin, 2024a, p. 3; Rubin, 2024b, p. 51).

### (2) Union-Intersection Inference

If a researcher makes a decision about a *union* alternative hypothesis (e.g., $H_{1,1} \cup H_{1,2} \cup H_{1,3} \cup \ldots H_{1,k}$) and corresponding *intersection* null hypothesis ($H_{0,1} \cap H_{0,2} \cap H_{0,3} \cap \ldots H_{0,k}$), then the Type I error rate for that decision is the familywise error rate of $1 - (1 - \alpha)^k$. In this case, the intersection null hypothesis can be rejected following *at least one* significant result (i.e., $p_{min} \leq \alpha$) among *any* of the $k$ tests ($p_1; H_{0,1}$, $p_2; H_{0,2}$, $p_3; H_{0,3}$, … $p_k; H_{0,k}$; e.g., García-Pérez, 2023, p. 15; Hochberg & Tamhane, 1987, p. 28; Parker & Weir, 2020, p. 563; Roy, 1953; Rubin, 2021b, pp. 10973–10975; Rubin, 2024b, pp. 51–52). Hence, there are *multiple* ($k$) opportunities to make a Type I error with regard to the intersection null hypothesis. Consequently, the familywise error rate for a decision about the intersection null hypothesis ($\alpha_{Intersection}$) is higher than the nominal error rate for tests of each of its constituent null hypotheses ($\alpha_{Constituent}$). To control $\alpha_{Intersection}$ at a conventional level (e.g., 0.05), $\alpha_{Constituent}$ must be adjusted below the conventional level (e.g., using a Dunn–Šidák correction, $\alpha_{Constituent} = 1 - [1 - \alpha_{Intersection}]^{1/k}$, or a Bonferroni correction, $\alpha_{Constituent} = \alpha_{Intersection}/k$).[3]

### (3) Intersection-Union Inference

If a researcher makes a decision about an *intersection* alternative hypothesis (e.g., $H_{1,1} \cap H_{1,2} \cap H_{1,3} \cap \ldots H_{1,k}$) and corresponding *union* null hypothesis ($H_{0,1} \cup H_{0,2} \cup H_{0,3} \cup \ldots H_{0,k}$), then the Type I error rate for that decision is the nominal $\alpha$ level. Here, the union null hypothesis can only be rejected following *all* significant results (i.e., $p_{max} \leq \alpha$) on *all* $k$ tests ($p_1; H_{0,1}$, $p_2; H_{0,2}$,

$p_3; H_{0,3}, \ldots p_k; H_{0,k}$). As with individual inferences, no alpha adjustment is required in this case because the researcher only has a *single* opportunity to erroneously reject the union null hypothesis (Berger, 1982, p. 295; Rubin, 2021b, p. 10976).

## Guiding Principles for Determining Type I Error Rates

The formal inference approach applies the following four guiding principles when determining relevant Type I error rates.

### (1) Type I Error Rates Refer to Random Sampling Error, not Systematic Error

A Type I error rate is represented by one or more critical regions in the tail areas of the specified sampling distribution of a test statistic given under a statistical null hypothesis (Neyman, 1950, p. 265). Hence, it is supposed to represent the probability of incorrectly rejecting a null hypothesis due to *random sampling* error *given* a specified sampling distribution. It is not supposed to quantify the probability of incorrectly rejecting a null hypothesis due to *systematic error* (Meehl, 1997, pp. 397, 401; Neyman & Pearson, 1928, p. 176; Pollard & Richardson, 1987, p. 162; Rubin, 2024a, p. 3, Confusion IV; Rubin, 2024b, p. 49). Systematic error can include (a) theoretical error (e.g., errors in theories, hypotheses, predictions, background knowledge, causal relations, constructs, etc.); (b) methodological error (e.g., errors in research design, sampling procedures, testing conditions, stimuli, manipulations, measures, controls, etc.); (c) data error (e.g., errors in data collection, entry, coding, cleaning, aggregation, transformation, etc.); or (d) analytical error (e.g., errors in statistical assumptions such as normality, linearity, or independence). A Type I error rate does not represent these types of systematic error. It represents a hypothetical, mathematical, subjunctive conditional probability that assumes that all test assumptions have been met, there is no systematic error, and the only remaining source of decision-making error is random sampling error, as represented by the sampling distribution (Cortina & Dunlap, 1997, pp. 166–167; Fisher, 1956, pp. 40, 44; Fisher, 1959, pp. 21–22; Kass, 2011, pp. 5–6, 7; Meehl, 1990, p. 111; Neyman, 1952, pp. 23, 57; Neyman, 1955, p. 17; Neyman & Pearson, 1928, pp. 176–177, 232; Rubin, 2024b, pp. 47–48).

Of course, systematic error can cause the substantive misinterpretation of a significant true positive result. For example, a theoretically uninteresting "crud" effect may be misinterpreted as a theoretically-predicted effect (Meehl, 1997, p. 402). However, it is important not to conflate substantive interpretational error with statistical Type I error (Greenland, 2017, p. 639; Meehl, 1990, p. 125; Rubin, 2017a, p. 274; Rubin, 2024b, pp. 50–51). The former is a substantive misinterpretation of a true positive result caused by systematic error. The latter is a statistical misclassification of a true negative (null) result caused by random error. A Type I error rate represents random error *given* a statistical model, not systematic error in the *choice* or *interpretation* of that model.

### (2) Type I Error Rates for Different Tests are Incommensurate with One Another

A Type I error rate is a hypothetical probability that is contingent on a specified sampling distribution under a statistical null hypothesis. It cannot be detached from its sampling distribution and treated as a standalone, unconditional probability that may become "inflated," "altered," "invalidated," or "vitiated" when moving from one test to another (Cortina & Dunlap, 1997, pp. 166–167; Pollard & Richardson, 1987; Weisberg, 2014, p. 289).

For example, the Type I error rate for one brand of pregnancy test may be lower than that for another brand, but it would be inappropriate to claim that the second test's error rate is an

"inflated" version of the first test's error rate or that the first test's error rate is "invalidated" in this context. Each error rate licenses a separate decision about a different statistical null hypothesis based on a different statistical model. In this respect, Type I error rates for different tests are independent and incommensurate with one another (García-Pérez, 2023, p. 15; Rubin, 2024a, p. 3).

### (3) Familywise Error Rates do not License Individual Inferences

The familywise error rate for the intersection null hypothesis $H_{0,1} \cap H_{0,2}$ only licenses a *union-intersection* inference about $H_{0,1} \cap H_{0,2}$ as a whole. Here, the constituent null hypotheses $H_{0,1}$ and $H_{0,2}$ are treated as being logically exchangeable with one another rather than as distinct individual hypotheses (García-Pérez, 2023, p. 2; Rubin, 2021b, pp. 10978–10982; Rubin, 2024a, p. 2). Consequently, during union-intersection testing, a significant result in relation to the constituent null hypothesis $H_{0,1}$ does not allow us to reject $H_{0,1}$ as an individual hypothesis. It only allows us to reject $H_{0,1} \cap H_{0,2}$ and conclude that either $H_{0,1}$ or $H_{0,2}$ or both are false.[4]

The same principle applies in the case of a one-way ANOVA (García-Pérez, 2023, p. 6). A significant omnibus result allows us to claim that there is a significant difference between at least one pair of means, but it does not allow us to identify which pair (Rubin, 2024a, p. 2).

Consequently, in the formal inference approach, a familywise error rate is not the probability of making "at least one Type I error" because this phrasing implies that the researcher is making multiple *individual* inferences (decisions) about multiple *individual* hypotheses, each licensed by its own nominal Type I error rate (Rubin, 2021b, pp. 10978–10983; Rubin, 2024a, p. 3, Confusion III; Rubin, 2024b, p. 59). Instead, as its name implies, the familywise error rate represents the probability of making an error about an entire *family* of hypotheses, not any specific individual hypothesis within that family. Hence, during union-intersection testing, the researcher only makes a *single* decision to either reject or fail to reject an intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2}$) on the basis of "at least one" significant result among their constituent tests (e.g., $p_1;H_{0,1}$ and $p_2;H_{0,2}$; see also García-Pérez, 2023, p. 5; Rubin, 2021b, p. 10982). The familywise error rate is then the probability of making a Type I error about this intersection null hypothesis rather than the probability of making at least one Type I error about multiple individual null hypotheses.

### (4) Type I Error Rates are Based on Formally Reported Inferences, not "Actual" Test Procedures

If a researcher's "actual" test procedure includes tests of $H_{0,1}$ and $H_{0,2}$, then it is possible for them to report statistical inferences about $H_{0,1}$, $H_{0,2}$, $H_{0,1} \cap H_{0,2}$, and/or $H_{0,1} \cup H_{0,2}$. Nothing in the description of the test procedure (i.e., "tests of $H_{0,1}$ and $H_{0,2}$") indicates which of these inferences will be reported. Of course, a researcher can privately intend and publicly plan to report one or more of these inferences. However, intentions and plans can change and, when they do, inferences and their associated error rates may change with them (for related points, see Weisberg, 2014, p. 289). For example, a researcher may plan to make an inference about $H_{0,1} \cap H_{0,2}$, which would require an alpha adjustment. However, they may then change their mind and decide to make two individual inferences instead, one about $H_{0,1}$ and one about $H_{0,2}$, neither of which require an alpha adjustment. Hence, from a formal inference perspective, it is formally reported statistical inferences, rather than "actual" test procedures and/or planned inferences, that determine relevant Type I error rates (Rubin, 2024b. pp. 59, 62). In particular, a formal statistical inference serves as the basis for the "rational reconstruction" of a hypothetical test procedure and its subjunctive

conditional error rate (Reichenbach, 1938, pp. 6–7). For example, an individual inference about $H_{0,2}$ is associated with a hypothetical test procedure that only refers to $H_{0,2}$ and not $H_{0,1}$.

This formal inference perspective is consistent with the view that "statistical analyses are merely thought experiments, informing us as to what would follow deductively under their assumptions" (Greenland, 2017, p. 640; see also Hennig, 2010, 2023; Kass, 2011, p. 7; Weisberg, 2014, p. 289). From this perspective, Type I error rates are contingent on *hypothetical* sampling distributions based on formally specified hypotheses, rather than "actual" sampling distributions based on "actual" test procedures (e.g., Cortina & Dunlap, 1997, pp. 166–167; see also Fisher, 1956, p. 44). Finally, the formal inference approach is consistent with Popper's methodological rule that scientific hypotheses must be "intersubjectively testable" so that, in principle, they can be tested by anyone (Popper, 1966, p. 218; Popper, 2002, p. 22). Unlike formally reported inferences, "actual" test procedures may fail Popper's requirement because they may be only partially reported (e.g., during *p*-hacking) and, consequently, they may not be implementable by others.

# 3. *p*-Hacking Does Not Inflate Type I Error Rates in the Formal Inference Approach

## The Basic Argument

To recap, the main claim is that *p*-hacking inflates Type I error rates in the error statistical approach but not in the formal inference approach. The basic argument for this claim is as follows.

In the error statistical approach, the "actual" familywise error rate is relevant during *p*-hacking because it covers both reported and unreported tests in the "actual" test procedure. In this approach, Type I error rate inflation occurs because the "actual" familywise error rate is higher than the nominal rate for the reported inference.

In contrast, in the formal inference approach, the "actual" familywise error rate is irrelevant because it only licenses an inference about an intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2}$), and, in practice, researchers who *p*-hack cannot report this inference because they do not disclose some of their tests (e.g., $p_1 > 0.05; H_{0,1}$). Instead, they usually report separate individual inferences about selectively reported individual null hypotheses (e.g., $p_2 \leq 0.05; H_{0,2}$).[5] Logically, the "actual" familywise error rate does not license these individual inferences. Rather, they are licensed by their own separate nominal error rates. *p*-hacking does not inflate these nominal error rates because inflation only occurs relative to the "actual" familywise error rate, and a comparison with the "actual" error rate is inappropriate because it is incommensurate and inconsistent with the reported individual inferences.

## An Illustration

To illustrate, imagine the following example of *p*-hacking: A researcher conducts a two-sided independent samples *t* test using a conventional alpha level of 0.05. They fail to find a significant result with regard to their first null hypothesis $H_{0,1}$: $t(326) = 1.88$, $p = 0.061$. *For that reason*, they remove an outlier from their sample and conduct a second test. The second test has a smaller degrees of freedom than the first, and it is formally specified by a different test procedure (i.e., outliers removed). Hence, it tests a different statistical null hypothesis, $H_{0,2}$. This time, the researcher finds a significant result, $t(325) = 2.16$, $p = 0.032$, which they then report without disclosing the nonsignificant result from their first test.

From an error statistical perspective, the researcher has conducted two significance tests ($p_1$;$H_{0,1}$ and $p_2$;$H_{0,2}$) and then selectively reported whichever test(s) yielded a significant result (Mayo, 1996, pp. 303–304, 348; Mayo, 2008, pp. 874–875; Mayo, 2018, pp. 274–275; Mayo & Cox, 2010, pp. 267–270). In this case, the "actual" sampling distribution is given under the "actual" test procedure's "global" or "universal" intersection null hypothesis, $H_{0,1} \cap H_{0,2}$ (Mayo, 2008, p. 875; Mayo, 2018, p. 276; Mayo & Cox, 2010, p. 269). Consequently, the "actual" error rate is the familywise error rate 0.098 (i.e., $1 - [1 - 0.05]^2$), which is inflated above the nominal error rate of 0.05.

In contrast, from a formal inference perspective, the relevant sampling distribution is given under $H_{0,2}$, not $H_{0,1} \cap H_{0,2}$, because the researcher only reported an *individual* inference about $H_{0,2}$, not a *union-intersection* inference about $H_{0,1} \cap H_{0,2}$. Consequently, the relevant test procedure is the formally reported procedure for $H_{0,2}$, in which any outliers are removed, and the relevant Type I error rate is the nominal error rate (Rubin, 2020, p. 381; Rubin, 2021b, p. 10991; Rubin, 2025, p. 10).

Importantly, from a formal inference perspective, it is not appropriate to compare the "actual" familywise error rate for a decision about $H_{0,1} \cap H_{0,2}$ with the nominal error rate for a decision about $H_{0,2}$ and argue that the former represents an "inflated" version of the latter. As discussed previously, these two error rates are incommensurate with one another because they refer to two separate decisions about two different statistical null hypotheses based on two different statistical models. Arguing that a test's Type I error rate has been inflated because another test has a larger error rate is like arguing that a person's height has been inflated because their friend is taller than them!

It is also logically inconsistent to use a familywise error rate to license an individual inference. Specifically, we would be making a *fallacy of division* or *ecological fallacy* in this case because we would be misapplying an aggregate-level union probability to an individual member of the aggregate (Selvin, 1958; Waller, 2018). In particular, the union probability of obtaining *at least one* significant result given $H_{0,1} \cap H_{0,2}$ (i.e., the familywise error rate) does not represent the individual probability of obtaining a significant result given $H_{0,2}$ alone (García-Pérez, 2023; Rubin, 2017a, pp. 271–272; Rubin, 2020, p. 380; Rubin, 2021a; Rubin, 2021b, pp. 10978–10983; Rubin, 2024a, p. 3; Rubin, 2024b, p. 51). To illustrate, consider 100 tests, each with $\alpha = 0.05$. In this case, there is a 0.99 union probability of finding a significant result that incorrectly rejects the intersection null hypothesis $H_{0,1} \cap H_{0,2} \cap H_{0,3} \ldots \cap H_{0,100}$ but only a 0.05 individual probability of finding a significant result that incorrectly rejects the individual null hypothesis $H_{0,1}$ (Rubin, 2021b, p. 10978).

The familywise error rate would only be relevant if $H_{0,2}$ was treated as a logically exchangeable constituent of $H_{0,1} \cap H_{0,2}$, rather than as a distinct individual hypothesis. In this case, the researcher would make a union-intersection inference about $H_{0,1} \cap H_{0,2}$ as a whole, rather than an individual inference about $H_{0,2}$ alone (García-Pérez, 2023, p. 2; Rubin, 2021b, p. 10981; Rubin, 2024a, p. 2). In particular, following at least one significant result among $p_1$;$H_{0,1}$ and $p_2$;$H_{0,2}$, the union-intersection inference would be that "either $H_{0,1}$ or $H_{0,2}$ or both are false" rather than "$H_{0,2}$ is false."

In summary, from an error statistical perspective, (a) the "actual" test procedure includes two significance tests: $p_1$;$H_{0,1}$ and $p_2$;$H_{0,2}$; (b) the "actual" sampling distribution is given under the intersection null hypothesis $H_{0,1} \cap H_{0,2}$; and so (c) the "actual" Type I error rate is the familywise error rate of 0.098 (i.e., $1 - [1 - 0.05]^2$), which is "inflated" relative to the nominal error rate of 0.05. In contrast, from a formal inference perspective, (a) the formally reported inference is an

individual inference about the individual null hypothesis $H_{0,2}$; (b) the relevant sampling distribution for this inference is given under $H_{0,2}$; and so (c) the relevant Type I error rate is $p_2$; $H_{0,2}$'s nominal error rate of 0.05, which is incommensurate with the familywise error rate for a decision about $H_{0,1} \cap H_{0,2}$. Figure 1 illustrates these differences.[6]

**Figure 1**
*p-Hacking in the Formal Inference and Error Statistical Approaches*
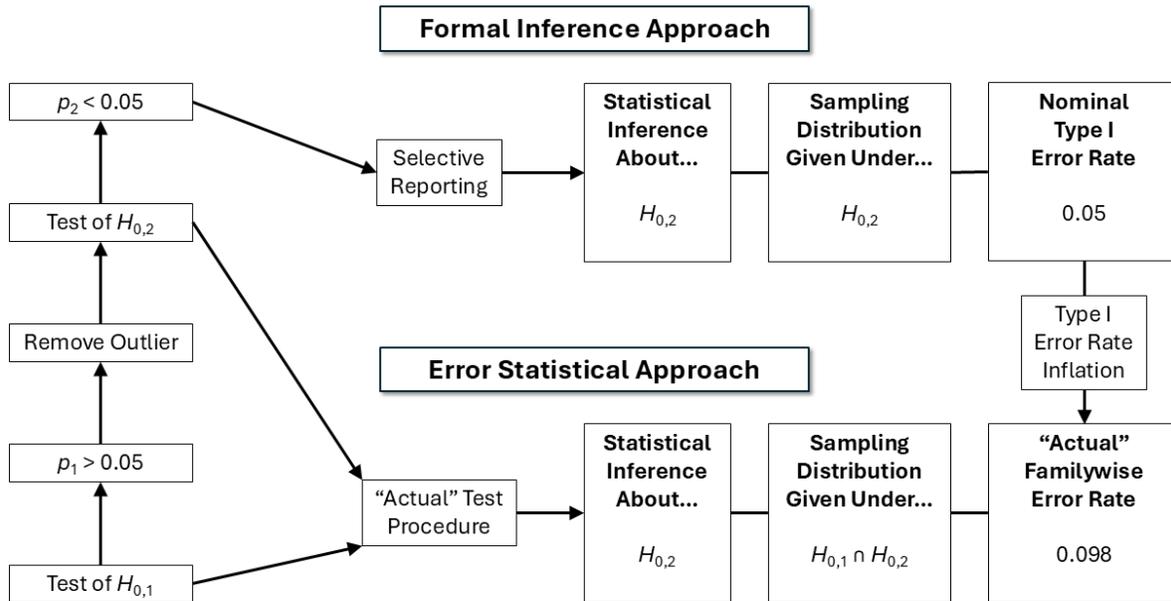


## A Closer Look

It is worth taking a closer look at the situation in Figure 1 in order to unpack some of its complexities. Contrary to the formal inference approach, the error statistical approach argues that the "actual" familywise error rate is relevant because we need to consider all of the hypotheses that could have been outputted by the "actual" test procedure (i.e., $H_{0,1}$ and $H_{0,2}$) "in order to assess the well-testedness of the one hypothesis [that it] … happened to have outputted" (i.e., $H_{0,2}$; Mayo, 2008, p. 876). As Mayo (2008) explained:

> The hypothetical error rates teach us about the test's capacities in the case at hand. Because at least one such impressive departure is common even if all are due to chance, the 'hunting expedition' has scarcely reassured us that it has done a good job of avoiding such a mistake in *this* case. The .05 'computed' *p*-value is invalidated when it comes to the 'actual' value. (p. 876, italics in original here and in all subsequent quotes)

Again, however, from a formal inference perspective, this reasoning commits the fallacy of division because it confounds union and individual probabilities. In the situation shown in Figure 1, the "actual" familywise error rate (a union probability) only informs us about the capacity of the "hunting expedition" to avoid an error in relation to the *intersection* null hypothesis $H_{0,1} \cap H_{0,2}$. It does not inform us about an *individual* test's capacity to avoid an error in relation to the *individual* null hypothesis $H_{0,2}$ (an individual probability), and it is $H_{0,2}$ that represents "the case

at hand" in Figure 1, not $H_{0,1} \cap H_{0,2}$. From this perspective, it is the individual test's nominal error rate that indicates its error-probing capacity with respect to $H_{0,2}$, and that capacity is not invalidated by the hunting expedition that was used to identify and selectively report the test. By analogy, if 20 athletes are tested for illicit drugs, then hunting for, finding, and selectively reporting a test that yields a positive result does not alter that test's capacity to avoid an error in relation to the athlete who took the test (see also Rubin, 2024b, p. 55).[7]

In theory, the researcher in Figure 1's *p*-hacking example could report a union-intersection inference based on the "actual" intersection null hypothesis $H_{0,1} \cap H_{0,2}$. For example, given different results, they could report that "there was a significant result when either including outliers, excluding outliers, or in both cases, $p_{\min} \leq 0.025$." In this case, the formal inference and error statistical approaches would agree that the familywise error rate is relevant because, logically, the relevant sampling distribution for this formally reported union-intersection inference is given under $H_{0,1} \cap H_{0,2}$ (Rubin, 2017b, p. 323; Rubin, 2021b, pp. 10973–10975; Rubin, 2024b, pp. 51–52). However, in practice, the researcher cannot report this inference because their *p*-hacking conceals the null result for $H_{0,1}$. Instead, they report an individual inference about $H_{0,2}$, and that inference is licensed by its own nominal Type I error rate.

Error statisticians might argue that we "ought" to audit the experimental testing context to try to detect any *p*-hacking and make this hidden part of the "actual" test procedure objective (e.g., Mayo, 2018, p. 49). However, from a formal inference perspective, publicly reporting previously undisclosed tests does not necessarily change the specified category of statistical inference. For example, the researcher might report both $p_1;H_{0,1}$ and $p_2;H_{0,2}$ but specify two *individual* inferences as follows:

> One outlier was identified at ±3.00 *SD*s from the sample mean. There was no significant result when this outlier was included in the analysis, $t(326) = 1.88$, $p = 0.061$. However, there was a significant result when the outlier was excluded, $t(325) = 2.16$, $p = 0.032$.

In this case, the Type I error rate for each individual statistical inference remains at its nominal level even if the second test was "actually" conducted *because* the first test yielded a null result (see also García-Pérez, 2023; Parker & Weir, 2020, pp. 563–564; Rubin, 2017a, pp. 271–272; Rubin, 2020, p. 380; Rubin, 2021a; Rubin, 2021b, pp. 10978–10979; Rubin, 2024a, p. 3, Confusion I; Rubin, 2024b, p. 51). Again, logically, the researcher would only need to refer to the familywise error rate if they formally reported a union-intersection inference about $H_{0,1} \cap H_{0,2}$. It would be a category mistake for them to use the familywise error rate to license individual inferences about $H_{0,1}$ and $H_{0,2}$, even if the result for $H_{0,1}$ inspired the researcher to test $H_{0,2}$ (Rubin & Donkin, 2024, pp. 2023–2024).

Mayo (1996) argued that the nominal error rate misrepresents "*what should be expected to happen … in subsequent experiments*" (p. 349, see also p. 310). This "error unreliability" argument assumes that the original experiment "actually" undertook a union-intersection test of $H_{0,1} \cap H_{0,2}$, rather than an individual test of $H_{0,2}$, and then evaluated the test result using the wrong (nominal) sampling distribution (Spanos & Mayo, 2015, p. 3546). However, in the formal inference approach, a significance test of $H_{0,2}$ can be treated as either (a) part of a union-intersection test of $H_{0,1} \cap H_{0,2}$ or (b) an individual test of the individual null hypothesis $H_{0,2}$, and it is the formally reported inference that determines which is the case. Hence, if an original experiment reports an individual inference about $H_{0,2}$, then, objectively, the significance test is treated as an individual test of $H_{0,2}$, and "what should be expected to happen" in hypothetical exact replications of this

individual test is licensed by the nominal Type I error rate for that test (see also Rubin, 2024b, p. 58).

Finally, it is important to appreciate that the error statistical scenario shown in the bottom half of Figure 1 represents a simplistic conceptualization of the "actual" test procedure because it ignores the indeterminate nature of *p*-hacking in a hypothetical long run of repeated random sampling. By definition, a researcher who *p*-hacks will continue testing until they obtain a desired result. Consequently, although the researcher in Figure 1 managed to obtain a significant result on their second attempt in the current sample, they may need many more attempts in other samples (e.g., by conducting additional tests that add covariates, use different outcome measures, etc.; Stefan & Schönbrodt, 2023). In this case, the number of tests (*k*) in hypothetical repetitions of the "actual" test procedure is unknown, and so we cannot compute the "actual" familywise error rate as $1 - (1 - \alpha)^k$ (Hochberg & Tamhane, 1987, p. 6; Mayo, 1996, pp. 313–314; Rubin, 2017b, p. 325; Rubin, 2021b, p. 10992; Rubin, 2024b, p. 56; see also Simmons et al., 2011, p. 1365).

From an error statistical perspective, this indeterminate testing problem invalidates significance testing during not only secretive *p*-hacking but also transparently reported result-dependent exploratory analyses (Rubin, 2017a; Rubin, 2021b, p. 10992; Rubin, 2024b, p. 56). Even so-called "honest hunting" for significance, which includes a correction for multiple testing, is untenable in an open-ended exploratory situation because transparency about the number of tests conducted in relation to one sample does not inform us about the number of tests that would be conducted in other samples (Mayo, 1996, pp. 311−314; Mayo, 2018, p. 275).

The formal inference approach avoids this problem by defining *k* relative to a closed, formally specified inference rather than an open-ended and indeterminate "actual" test procedure (Rubin, 2024b, p. 64). For example, $k = 1$ for an individual inference about $H_{0,1}$, and $k = 2$ for a union-intersection inference about $H_{0,1} \cap H_{0,2}$, even if these two inferences have been selectively reported from a larger set of unreported inferences. Hence, from a formal inference perspective, significance testing retains its value in unplanned exploratory analyses (Rubin, 2017a, 2020, 2022, 2024b; Rubin & Donkin, 2024).

# 4. Areas of Agreement and Disagreement

## Areas of Agreement

Despite their contrasting views on *p*-hacking, there are several areas of agreement between the error statistical and formal inference approaches that are worth highlighting. First, they both acknowledge that their diverging views reflect different philosophies of science (e.g., Mayo, 1996, p. 255; Mayo & Cox, 2010, p. 267; Rubin, 2024b, pp. 46–47; Rubin, 2025). As Mayo and Cox (2010) explained:

> The general issue is whether the evidential bearing of data *y* on an inference or hypothesis $H_0$ is altered when $H_0$ has been either constructed or selected for testing in such a way as to result in a specific observed relation between $H_0$ and *y*, whether that is agreement or disagreement. Those who favor logical approaches to confirmation say no (e.g., Mill 1888, Keynes, 1921), whereas those closer to an error-statistical conception say yes (Whewell (1847), Pierce (1931–5)). (p. 267)

Error statisticians also acknowledge the logical and mathematical arguments that support the formal inference approach, although they find these arguments wanting. For example, Mayo and Cox (2010) explained that, "to the extent that *p* is viewed as an aspect of the logical or

mathematical relation between the data and the probability model, … preliminary choices [based on an inspection of the data] are irrelevant" (p. 266). Similarly, Mayo (1996) explained that, "if confirmation is strictly a logical function between evidence (or statements of evidence) and hypotheses, when or how hypotheses are constructed will be irrelevant" (p. 255).

Finally, the two approaches agree that a researcher's personal biases may affect the specification of a significance test without invalidating its Type I error rate. As Mayo (1996) explained:

> It does not matter that test specifications might reflect the beliefs, biases, or hopes of the researcher. Perhaps the reason for selecting an insensitive test is your personal desire to find no increased risk, or perhaps it is due to economic or ethical factors. Those factors are entirely irrelevant to scrutinizing what the data do and do not say. They pose no obstacle to my scrutinizing any claims you might make based on the tests, nor to my criticizing your choice of test as inappropriate for given learning goals. (p. 409; see also pp. 148, 263)

> The latitude that exists in the choice of test does not prevent the determination of what a given result does and does not say. The error probabilistic properties of a test procedure — *however that test was chosen* — allows for an objective interpretation of the results. (p. 406)

## Areas of Disagreement

Despite agreeing that personally biased test specification does not prevent an objective interpretation of test results, the two approaches disagree about the impact of undisclosed result-dependent selective reporting (i.e., *p*-hacking). According to the error statistical approach, *p*-hacking invalidates significance testing because it represents an unrecognized part of the relevant ("actual") test procedure. In contrast, in the formal inference approach, *p*-hacking is not part of the relevant (formally reported) test procedure. Consequently, and consistent with its view on personally biased test specification, the formal inference approach assumes that *p*-hacking does not prevent either (a) public scrutiny of the formally reported rationale for a specified test (Rubin, 2022, pp. 539–542; Rubin & Donkin, 2024, pp. 2024, 2036–2037) or (b) an objective interpretation of the test's results relative to its nominal Type I error rate.

From a broader perspective, the two approaches disagree about where to draw the line between the *context of discovery* and the *context of justification* (Reichenbach, 1938, pp. 6–7). The error statistical approach argues that the construction and selection of hypotheses and data occur in the context of justification. In contrast, the formal inference approach follows Popper's (1983) view that "the factual, psychological, and historical question, 'How do we come by our theories?' [or statistical hypotheses], though it may be fascinating, is irrelevant to the logical, methodological, and epistemological question of validity" (p. 36; see also Popper, 2002, pp. 7–8, 22). From this perspective, it does not matter that statistical hypotheses are secretly chosen because their test results are psychologically "desired" by researchers (e.g., Nagy et al., 2025, p. 2; Simmons et al., 2011, p. 1359; Stefan & Schönbrodt, 2023, p. 2). What matters is a public critical evaluation of the objective, formally reported, test procedures and inferences.

Relatedly, the error statistical and formal inference approaches have different responses to the concern that the specification of an "actual" test procedure may depend on psychological "intentions locked up inside the scientist's head" (Mayo & Spanos, 2011, p. 186; see also Mayo, 1996, pp. 346–350; Mayo, 2008, pp. 860–861). To address this concern, the error statistical

approach argues that we should audit the scientist's experimental testing context in order to "pick up on" relevant information regarding "the construction and selection of both hypotheses and data" that can then be use to specify the "actual" test procedure (Mayo, 2018, p. 49). In contrast, the formal inference approach argues that the informal testing context and "actual" test procedure are irrelevant, and we should only refer to the scientist's objective, formally reported, but nonetheless hypothetical test procedure (Rubin, 2025, p. 10).

Mayo (1996) recognized these diverging views when defining the "actual" test procedure in the context of an example in which a researcher conducted multiple tests and then searched for a significant result to report:

> It may be objected that … I am taking sides in favor of one description of the "actual" test procedure — one that takes into account the fact that searching has occurred. I am, but maintain that this aspect of the procedure cannot be ignored given the aim of the statistical significance test chosen. Remember, I am distinguishing the appropriateness of the test chosen (for a given inquiry) from the error probabilities, *given* that that test is chosen. (p. 304)

Mayo's (1996) position makes sense from an error statistical perspective because ignoring the undisclosed searching for significance misspecifies the "actual" test procedure. However, in the formal inference approach, "the aim of the statistical test chosen" is to tentatively rule out random sampling error "*given* that that test is chosen" and "*however that test was chosen*" (Mayo, 1996, pp. 304, 406). From this perspective, it is legitimate to define the test procedure solely in terms of its formal, publicly reported specification and independent from any personally biased, result-dependent selection process that influenced that specification (see also Mayo, 1996, p. 255; Mayo & Cox, 2010, p. 266).

Mayo and Cox (2010) disagree, arguing that, "by allowing the result to influence the choice of specification, one is altering the procedure giving rise to the *p*-value" (p. 271). Again, they are correct from an error statistical perspective because an error statistical test includes the process by which its hypothesis and data were constructed and selected (Mayo, 1996, pp. 206–207, 298). Consequently, allowing a test's result to influence this process will alter the "actual" test procedure. However, from a formal inference perspective, a significance test and its *p*-value are only conditioned on the formally reported test procedure and not on the "factual, psychological, and historical" process that led up to that procedure (Popper, 1983, p. 36). Consequently, changing the process by which a test procedure is specified does not alter the procedure itself.

Mayo (2018) claimed that "selection effects alter the outcomes in the sample space, showing up in altered error probabilities" (p. 286). Again, this claim is correct if we follow the error statistical approach and extend the sample space to include outcomes that would arise during hypothetical repetitions of the selection process. However, it does not apply in the formal inference approach, which limits the relevant sample space to outcomes under the selected statistical null hypothesis. For example, a biasing selection effect will not alter the Type I error rate for an inference about $H_{0,13}$ when that inference is limited to potential outcomes in $H_{0,13}$'s sample space, even if $p_{13};H_{0,13}$ has been *p*-hacked from a set of unreported tests $p_1;H_{0,1}, \ldots p_k;H_{0,k}$ (Rubin, 2024b, p. 55).

Spanos and Mayo (2015) argued that "the discrepancy between actual and nominal error probabilities stems from evaluating the nominal error probabilities using the wrong sampling distribution" (p. 3546; see also Figure 1). Certainly, from an error statistical perspective, the

sampling distribution under $H_{0,13}$ does not reflect the "actual" test procedure when $p_{13};H_{0,13}$ has been *p*-hacked from $p_1;H_{0,1}$, $p_2;H_{0,2}$, $p_3;H_{0,3}$, … $p_k;H_{0,k}$. However, from a formal inference perspective, the sampling distribution under the intersection null hypothesis $H_{0,1} \cap H_{0,2} \cap H_{0,3}$ … $\cap H_{0,k}$ is the "wrong" distribution because it is incommensurate and inconsistent with the researcher's formally reported inference, which is only about the individual hypothesis $H_{0,13}$ (Rubin, 2024a, p. 3; Rubin, 2025, pp. 13–14).

Finally, Mayo (2018) returns to the issue of well-testedness or severity:

> For the severe tester, outputting $H_{13}$, ignoring the non-significant others, renders $H_{13}$ poorly tested. You might say, but look there's the hypothesis $H_{13}$, and data $x$ — shouldn't it speak for itself? No. That's just the evidential-relationship or logicist in you coming out. (p. 275)

Again, however, in the formal inference approach, the well-testedness of $H_{0,13}$ depends on the individual test of $H_{0,13}$, not the union-intersection test of $H_{0,1} \cap H_{0,2} \cap H_{0,3}$ … $\cap H_{0,k}$. In this respect, formal inferentialists are not afraid to let their inner logicist come out! Indeed, they would view it as illogical to use the "actual" familywise error rate to license an individual inference about $H_{0,13}$ because this approach commits the fallacy of division (Selvin, 1958; Waller, 2018; for examples of this fallacy in real research studies, see Rubin, 2024a, pp. 5–6).

In summary, disagreements between the error statistical and formal inference approaches stem from their different conceptualizations of test procedures, the context of justification, sample spaces, and sampling distributions. The error statistical approach assumes that *p*-hacking (a) constitutes part of the "actual" test procedure, (b) occurs in the context of justification, and (c) alters the relevant sample space and sampling distribution. In contrast, the formal inference approach assumes that *p*-hacking (a) is not part of the selected, formally reported test procedure, (b) occurs in the context of discovery and, consequently, (c) does not alter the relevant hypothetical sample space or sampling distribution (Rubin, 2025, pp. 4–5, 9–10).

# 5. The Minimal Severity Requirement

According to the error statistical approach's minimal severity requirement, it is important to adequately check a test's assumptions in order to provisionally rule out systematic error, because systematic error can cause a significant discrepancy between "actual" and nominal error probabilities that invalidates nominal error probabilities (Mayo & Spanos, 2011, p. 190; Spanos, 2010, p. 219; Spanos & Mayo, 2015, pp. 3541, 3543). In the formal inference approach, however, nominal Type I error rates are not invalidated by systematic error because they refer to a *hypothetical* situation in which there is no systematic error, all test assumptions are met, and the only source of decision-making error is random sampling error (Fisher, 1956, pp. 40, 44, 77–78, 82; Fisher, 1959, pp. 21–22; Kass, 2011, pp. 5–6, 7; Meehl, 1990, p. 111; Neyman, 1955, p. 17; Neyman & Pearson, 1928, pp. 177, 232; Rubin, 2024b, p. 49; see also Popper, 1983, p. 313). Hence, it is not necessary to meet the minimal severity requirement in order to obtain valid Type I error rates. This point can be illustrated with respect to (a) implicatory assumptions and (b) the Texas sharpshooter fallacy.

## Implicatory Assumptions

In the formal inference approach, a Type I error rate refers to a hypothetical situation in which random sampling error is the only source of decision-making error. This hypothetical situation can be described as an "*implicationary* or *i-assumption*" that is made during testing in

order to draw out certain logical implications (Mayo, 2018, pp. 109, 167). According to Mayo (2018), "the howler [mistake] occurs when a test hypothesis that serves merely as an i-assumption is purported to be an actual assumption, needed for the inference to go through" (p. 167; see also Cortina & Dunlap, 1997, pp. 166−167). The formal inference approach agrees and applies the same reasoning to not only the test hypothesis but also the test assumptions (see also Hennig, 2010, p. 47; Hennig, 2023, pp. 24, 38). Hence, the formal inference approach does not require a test's "actual" assumptions to be checked and affirmed as "approximately true" in order for a statistical inference to go through (Mayo, 2011, p. 96; Mayo & Spanos, 2011, p. 189; cf. Hennig, 2023, p. 38). Instead, during testing, all test assumptions are provisionally accepted as i-assumptions. In this case, failure to probe for systematic error in a test's "actual" assumptions does not invalidate that test's nominal error rate because the associated inferential argument is based on unproblematic i-assumptions.

The relation between "actual" and implicatory assumptions is similar to that between the soundness and validity of a logical argument. For example, consider the argument that "all cats are gray, and *x* is a cat; therefore, *x* is gray." This argument is logically valid, even though it is unsound because its major premise is not "approximately true" — not all cats are gray; not even most cats are gray (Fisher, 1959, pp. 22−23; Mayo, 2018, p. 60). Furthermore, the argument remains valid even though other major premises are available that *are* approximately true (e.g., "all cats have whiskers"). Similarly, in the formal inference approach, it remains valid to use a nominal Type I error rate to license a statistical inference even when the associated test's "actual" assumptions are not approximately true. In other words, it remains valid to make the suppositional, subjunctive conditional argument that, *if* a test's assumptions *were* true and random sampling error was the only source of decision-making error, *then* the nominal Type I error rate would represent the maximum frequency with which the null hypothesis would be incorrectly rejected (Neyman, 1950, p. 289; Neyman, 1955, p. 17; see also Cortina & Dunlap, 1997, pp. 166–167; Fisher, 1959, pp. 21–22; Hennig, 2023, pp. 21, 23–24; Kass, 2011, pp. 5–6, 7; Mayo, 2018, pp. 426–427). In this case, the nominal Type I error rate represents a *hypothetical* error rate rather than an "actual" error rate, and "we speak of what *would* be inferred if our assumptions *were* to hold" (Kass, 2011, p. 7; see also Fisher, 1956, pp. 44, 77).[8]

The use of i-assumptions during hypothesis testing is also consistent with Popper, who argued that a test's background assumptions (initial conditions and auxiliary hypotheses) must be accepted as "unproblematic" during the test in order to force the test hypothesis into logical isolation (Popper, 1983, pp. 186, 244; Popper, 1994, p. 160; Popper, 2002, pp. 197, 260; see also Lakatos, 1978, pp. 32, 42; Rubin, 2025, p. 12, Footnote 7). Of course, a test's "actual" assumptions may be challenged before or after the test takes place (Lakatos, 1978, p. 158; Popper, 1994, p. 160), and if they are judged to be inadequate, then an alternative test may be considered (Popper, 2002, p. 86). However, in the formal inference approach, the identification of a more adequate test does not invalidate a *statistical* inference based on the original (inadequate) test because the inferential argument for that test is based on hypothetical i-assumptions, not real-world "actual" assumptions. "2 + 2 = 4," and this mathematical equation remains valid even if there has been a data entry error and one of the "2s" is later revealed to be a "3."

**The Texas Sharpshooter Fallacy**

Mayo also used the Texas sharpshooter fallacy to illustrate the error statistical concern about violating the minimal severity requirement (Mayo, 1996, pp. 201–203; Mayo, 2018, pp. 5, 19, 276–277; Mayo & Cox, 2010, p. 271). In this scenario, an individual with poor shooting skills

fires several shots at a barn wall and then paints a target around a random cluster of their bullet holes. Later on, they show these "bullseyes" to their friend as evidence of their "excellent shooting skills." The minimal severity requirement is not met in this scenario because there has been no attempt to rule out the sharpshooter's fraudulent behind-the-scenes behavior as one of the ways in which their claim may be false (Mayo, 2018, p. 19). Consequently, we are left with "bad evidence, no test (BENT)" (Mayo, 2018, p. 5).

Again, however, from a formal inference perspective, it is not necessary to meet the minimal severity requirement in order to report valid hypothetical probability statements. For example, it remains valid for the sharpshooter to compute the probability that their shots would hit the painted target in repetitions of their formally reported, hypothetical, test procedure (i.e., the procedure that they reported to their friend; see also Mayo, 2018, p. 19). In this case, the relevant probability is a formal mathematical probability about the occurrence of a hypothetical event based on chance alone, and so it is not invalidated by the presence of fraud in the real world (Fisher, 1956, pp. 40, 44, 77; Fisher, 1959; Neyman, 1955, p. 17; Neyman & Pearson, 1928, pp. 176, 232; cf. Mayo, 2018, p. 19).

To be clear, I am not arguing that the sharpshooter's fraudulent behavior is unproblematic. It certainly invalidates the *substantive* (nonstatistical) inference about their "excellent shooting skills." I am only arguing that, during significance testing, Type I error rates license hypothetical, test-specific, *statistical* inferences given random sampling error alone, not real-world, general, *substantive* inferences given systematic error (Meehl, 1990, p. 111; Meehl, 1997, p. 401; Neyman, 1950, p. 290). Consequently, we should not expect Type I error rates to be "altered," "invalidated," or "vitiated" by the presence of systematic error in the real world, including error caused by fraud.

By analogy, we would not expect a weighing scale's random measurement error to be altered or invalidated when a person lies about their weight or when the scale systematically underestimates weight. These systematic errors will certainly invalidate substantive inferences about the person's weight, but they will not invalidate statistical inferences based on the scale's random measurement error.

# 6. The Nondisclosure of Null Results

*p*-hacking results in the nondisclosure of null (nonsignificant) results. Again, this nondisclosure only inflates Type I error rates in the error statistical approach. Nonetheless, it may also be problematic in the formal inference approach, depending on how we conceptualize null results and their impact on substantive inferences.

To illustrate, consider a researcher who reports the substantive inference that "eating jelly beans causes acne" (Munroe, 2011). Imagine that they base this substantive inference on a single statistical inference: "Participants who ate green jelly beans had significantly more acne than those in a control group, $t(326) = 2.63$, $p = 0.009$." However, unbeknownst to readers, the researcher does not report the results of 19 other tests that found no significant results using 19 other colors of jelly beans ($p$s > 0.05).

From an error statistical perspective, the "actual" familywise error rate for the statistical inference (0.642) is inflated above the conventional nominal error rate (0.05) because the "actual" test procedure includes 20 tests, not 1 test. In contrast, from a formal inference perspective, the nominal Type I error rate of 0.05 is valid because the reported statistical inference is an individual inference about green jelly beans, not a union-intersection inference about green, red, blue, and other colors of jelly beans (for a discussion, see Rubin, 2021b, pp. 10978–10983). But does the researcher's failure to report the 19 null results bias their *substantive* inference that "eating jelly

beans causes acne"? The answer depends on how we conceptualize null results and how they relate to the substantive inference.

First, null results may be regarded as providing *evidence of absence* because we tentatively "accept" null hypotheses following severe, high-powered tests (Neyman, 1950, pp. 259–260; e.g., Mayo, 2018, p. 150; Mayo & Cox, 2010, p. 256; Mayo & Spanos, 2006, p. 339; Mayo & Spanos, 2011, p. 177). In this case, nondisclosed null results may bias substantive inferences. For example, in the current scenario, 19 results indicate that jelly beans *do not* cause acne, and only one result indicates that they *do*! Hence, failing to report the 19 null results is problematic because it hides relevant evidence and biases the evidential support for the substantive inference.

Second, null results may be regarded as the *absence of evidence* (Aczel et al., 2018; Altman & Bland, 1995; Murphy et al., 2025; Rubin & Donkin, 2024, p. 2036) because we merely "fail to reject" null hypotheses following nonsevere, low-powered, or Fisherian tests (Fisher, 1956, p. 45; Mayo & Spanos, 2006, pp. 338–339; Mayo & Spanos, 2011, p. 176). Hence, outside of a meta-analysis, null results do not impact substantive inferences because they do not provide any evidence one way or the other. For example, in the current scenario, the single significant result for green jelly beans supports the substantive claim that "eating jelly beans causes acne," whereas the 19 null results do not count as evidence either for or against this claim and so can be ignored.[9] From this perspective, it is more problematic for researchers to omit tests that yield *significant* results that *contradict* their substantive inference. For example, it would be problematic if the researcher failed to report a result showing that eating *red* jelly beans significantly *reduced* acne relative to a control group. In this case, they would be omitting relevant evidence that contradicted the substantive claim that "eating jelly beans causes acne" (Rubin, 2020, p. 384).

Finally, even if null results are regarded as providing evidence of absence, their nondisclosure will only bias a substantive inference when they are *theoretically relevant* to that inference (Rubin, 2020, p. 378; Rubin, 2022, p. 548). For example, failing to report the severe, high-powered, statistical inference that "there was no significant difference in acne between participants who ate M&Ms and those in a control group, $t(326) = 0.33$, $p = 0.742$" will not bias the substantive claim that "eating jelly beans causes acne" because M&Ms are not jelly beans (Rubin, 2021b, p. 10988). Hence, it is important to distinguish between *biased* selective reporting, which is problematic, and *unbiased* selective reporting, which is an essential part of effective scientific communication (Rubin, 2020, p. 383). A theory-centric critical rational discussion among researchers is required to make this distinction (Popper, 1966, pp. 218–219, 229–231; Popper, 1994, pp. 93, 159–16; Rubin, 2025, p. 11).

In summary, nondisclosed null results only bias substantive inferences when they are regarded as being both (a) evidence of absence and (b) theoretically relevant. How can we identify these evidentially and theoretically important null results? Preregistration represents one approach. However, preregistration cannot identify nondisclosed null results in the case of either exploratory analyses or unreported deviations from preregistered analyses. For example, a researcher may try out two deviations from a preregistered analysis protocol and then only report the deviation that yields a significant result. In this case, cross-referencing between the final research report and the preregistered protocol will not help to identify the undisclosed null result.

A more comprehensive approach to identifying theoretically important null results is to engage in a critical rational discussion of relevant theory, evidence, and background knowledge (Popper, 1966, 1994). As Kerr (1998) asked, if an original hypothesis is not reported, but it "had a sufficient rationale (theoretical, empirical, or even intuitive) to recommend itself to one researcher, why would it not also occur to others?" (p. 208; see also Rubin, 2020, p. 378; Rubin,

2022, p. 548). For example, in the current scenario, we might reasonably expect reviewers and other readers to ask why the researcher only used *green* jelly beans to test the substantive hypothesis that "eating jelly beans causes acne" and whether they tested any other colors of jelly beans (Wells & Windschitl, 1999). Furthermore, in the formal inference approach, researchers' anticipation of this critical rational discussion of their work helps to constrain their "researcher degrees of freedom" (Simmons et al., 2011, p. 1359) and limit an "anything goes" approach (Hennig, 2023, pp. 17, 19).[10]

# 7. Conclusions and Implications

## Conclusions

I have argued that *p*-hacking inflates Type I error rates in the error statistical approach but not in the formal inference approach. In the error statistical approach, the "actual" test procedure determines the relevant error rate. In contrast, in the formal inference approach, the formally reported statistical inference determines the relevant error rate. *p*-hacking is part of the "actual" test procedure, but it is not part of the procedure implied by the formally reported inference. Consequently, *p*-hacking inflates the relevant error rate in the error statistical approach but not in the formal inference approach.

Importantly, from a formal inference perspective, it is inappropriate to use the "actual" familywise error rate to license inferences about selectively reported hypotheses following *p*-hacking. Logically, the "actual" familywise error rate only licenses an inference about the entire family of reported and unreported hypotheses that are tested in the "actual" test procedure. It does not license inferences about selectively reported hypotheses, which are licensed by their own nominal error rates. Certainly, *p*-hacking influences the *selection* of reported significance tests. However, this selection process is not *part of* those tests, and so *p*-hacking does not influence the associated nominal Type I error rates. Hence, in the formal inference approach, "computed" nominal Type I error rates can be taken at face value, barring any logical or mathematical errors (Neyman, 1950, p. 289; Rubin, 2024b, p. 53; Rubin, 2025, p. 14).

In the error statistical approach, failing to meet the minimal severity requirement leaves us with "bad evidence, no test (BENT)" (Mayo, 2018, p. 5). In contrast, in the formal inference approach, failing to rule out ways in which a substantive inference could be false may result in a "bad" *substantive* inference without invalidating a corresponding *statistical* inference, which is based on tentatively accepted i-assumptions. Of course, a researcher's "actual" test assumptions may be judged to be inadequate. In this case, they should use a different test with more adequate assumptions. However, doing so only has the potential to reduce systematic error. It does not result in a more valid Type I error rate because Type I error rates do not represent systematic error in a test's selection, specification, or assumptions. They only represent random error *given* a test's selection, specification, and assumptions.

Finally, nondisclosed null results inflate Type I error rates in the error statistical approach but not in the formal inference approach. Nonetheless, nondisclosed null results may be problematic in the formal inference approach if they are regarded as being both (a) evidence of absence and (b) theoretically relevant. In this case, although unreported null results do not inflate Type I error rates, they may bias substantive inferences. However, this substantive bias can be identified and addressed through a theory-centric critical rational discussion among peers (Popper, 1966, pp. 218–219, 229–231; Popper, 1994, pp. 93, 159–160; Rubin, 2025, pp. 5, 11).

## Implications

The contrast between the error statistical and formal inference approaches has implications for the way in which we conceptualize *p*-hacking. *p*-hacking is often described as a "questionable research practice" (John et al., 2012). The current analysis supports this equivocal description as opposed to stricter characterizations that equate *p*-hacking with fraud and unethical research behavior (e.g., Miller et al., 2025; Pickett & Roche, 2018). In particular, *p*-hacking may be viewed as being more or less problematic for Type I error rates depending on one's philosophy of significance testing. Accordingly, it would be helpful for discussions about the potential dangers of *p*-hacking to be more clearly situated within relevant philosophies of significance testing in order to reveal otherwise implicit assumptions.

A similar recommendation applies to demonstrations and simulations showing that *p*-hacking inflates Type I error rates. This work tends to adopt an error statistical perspective by contrasting nominal and "actual" error rates (e.g., Simmons et al., 2011, p. 1359). However, it should be acknowledged that, during *p*-hacking, (a) researchers cannot report union-intersection inferences about "actual" intersection null hypotheses, (b) "actual" familywise error rates cannot be computed due to the indeterminate number of tests in a hypothetical long run of sampling, and, even if they could be computed, (c) "actual" familywise error rates do not license researchers' selectively reported individual inferences (Rubin, 2024b, pp. 58, 63).

The contrast between the error statistical and formal inference approaches also has implications for interventions that are intended to reduce *p*-hacking. In particular, from an error statistical perspective, preregistration provides a useful method of limiting *p*-hacking, Type I error rate inflation, and the nondisclosure of null results (Lakens, 2019; Mayo, 2018, pp. 106–107, 439). However, from a formal inference perspective, *p*-hacking does not alter relevant (nominal) Type I error rates (Rubin, 2024b, 2025). In addition, nondisclosed null results may (a) represent the absence of evidence, (b) be theoretically irrelevant, and (c) be identified following a critical rational discussion of relevant theory and background knowledge (Rubin, 2020, p. 378; Rubin, 2022, p. 548; Rubin & Donkin, 2024, p. 2036). Hence, the formal inference approach does not provide a rationale for preregistration based on either Type I error rate inflation or the nondisclosure of null results.

Turning to the replication crisis, error statisticians have argued that "inflated error rates [are] at the heart of obstacles to replication" (Mayo, 2025, p. 1) and that "much of the handwringing about irreproducibility is the result of wearing blinders as to the construction and selection of both hypotheses and data" (Mayo, 2018, p. 49). In contrast, from a formal inference perspective, researchers who *p*-hack do not report statistical inferences that have inflated Type I error rates because they do not report union-intersection inferences about "actual" intersection null hypotheses (Rubin, 2024b). Instead, they selectively report individual inferences that have valid nominal error rates.

To be clear, the formal inference approach attributes replication failures to *both* statistical Type I errors *and* substantive interpretational errors. However, it argues that, by definition, Type I error rates are fixed at their nominal levels. In contrast, unrecognised interpretational errors may occur more frequently than expected, which may help to explain lower than expected replication rates (i.e., the replication crisis; Rubin, 2024b, pp. 63–64).

Of course, it is important to undertake rigorous auditing in original studies to try to identify interpretational errors (Mayo, 2018, pp. 100, 441–442). However, auditing can only examine *known unknowns* — the potential interpretational errors that researchers have thought to investigate. By definition, even the most rigorous auditing cannot identify *unknown unknowns* —

the potential interpretational errors that researchers have not yet considered. Unlike known unknowns, unknown unknowns cannot be tentatively ruled out in original studies, and they may go on to represent hidden moderators that cause replication failures (Rubin, 2024b, pp. 63–64; for examples, see Firestein, 2016).

Hence, from a formal inference perspective, the replication crisis is not the result of *p*-hacking inflating Type I error rates. Certainly, Type I errors will result in *some* replication failures, but no more than expected based on the nominal conventional alpha level. The replication crisis is also not the result of poorly audited test assumptions. More rigorous auditing will help to tentatively rule out some interpretational errors in original studies. However, by definition, it will be unable to identify unrecognized interpretational errors (unknown unknowns) that go on to cause replication failures via hidden moderators.

Instead, the replication crisis is a result of our underestimation of the extent and impact of unrecognized interpretational errors. We end up being surprised by the extent of our own unanticipated ignorance! From this perspective, the solution to the replication crisis will not be found in new statistical or methodological approaches. Instead, it requires a more fundamental epistemological shift in our research culture that produces a better acknowledgment "of the finitude and fallibility of our knowledge, and of the infinity of our ignorance" (Popper, 1994, p. 123).

# **Endnotes**

1. Following the error statistical approach, a semicolon ";" is used to indicate that a *p*-value is computed *assuming* that a specific null hypothesis is true (e.g., "$p_1$;$H_{0,1}$"; Mayo & Spanos, 2006, p. 331; Mayo & Spanos, 2011, p. 169).

2. The word *actual* is placed in scare quotes throughout because, from a formal inference perspective, the "actual," inferentially relevant, test procedure and Type I error rate are both hypothetical and formally specified (Rubin, 2025, pp. 13–14).

3. Another method of correcting for multiple testing is based on the false discovery rate (FDR; Benjamini & Hochberg, 1995, p. 290). The familywise error rate is the probability of rejecting at least one constituent null hypothesis (e.g., $H_{0,1}$) in an intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2}$) when all are true. In contrast, the FDR is the expected proportion of incorrectly rejected null hypotheses among *all* (true *and* false) rejected null hypotheses. Hence, unlike the familywise error rate, the FDR assumes that some of the constituent null hypotheses in an intersection null hypothesis (e.g., $H_{0,1}$) may be *correctly* rejected because they are *false* (Rubin, 2021b, p. 10975). Consequently, from a formal inference perspective, the FDR does not represent a Type I error rate because Type I error rates are computed under the assumption that the associated null hypothesis is true (Neyman, 1952, p. 57), and the FDR assumes that the intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2}$) may be false because one or more of its constituent hypotheses may be false (e.g., $H_{0,1}$).

4. In response to Rubin (2021b), Goeman (2022) argued that strong control over the familywise error rate (e.g., via a Bonferroni correction) allows us to make inferences about not only (a) a superordinate intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2} \cap H_{0,3}$), but also (b) subordinate intersection null hypotheses (i.e., $H_{0,1} \cap H_{0,2}$, $H_{0,2} \cap H_{0,3}$, and $H_{0,1} \cap H_{0,3}$) and (c) individual null hypotheses (e.g., $H_{0,1}$, $H_{0,2}$, and $H_{0,3}$). However, Goeman's argument confuses (a) constituent null hypotheses that are treated as being logically exchangeable parts of the same superordinate intersection null hypothesis with (b) subordinate intersection and individual null hypotheses that are treated as being logically distinct and nonexchangeable with one another

(García-Pérez, 2023, p. 2; Rubin, 2021b, pp. 10978–10982; Rubin, 2024a, p. 2). For example, the familywise error rate for a union-intersection inference about $H_{0,1} \cap H_{0,2} \cap H_{0,3}$ only licenses a decision about $H_{0,1} \cap H_{0,2} \cap H_{0,3}$. It would be a fallacy of division (Waller, 2018) to argue that it also licenses decisions about subordinate intersection null hypotheses, such as $H_{0,1} \cap H_{0,2}$, and individual null hypotheses, such as $H_{0,1}$. Furthermore, in the formal inference approach, researchers should not be interested in the error rate for $H_{0,1} \cap H_{0,2} \cap H_{0,3}$ unless they report a union-intersection inference about this hypothesis. Customized, inference-specific error control at the level of formally reported intersection and/or individual hypotheses provides a more logically consistent and statistically powerful approach, and it makes strong error control over superordinate intersection null hypotheses irrelevant (see also Rubin, 2021b, pp. 10990–10992).

5. Throughout this article, I assume that researchers tend to report individual inferences rather than union-intersection or intersection-union inferences because individual inferences are the most common type of statistical inference (García-Pérez, 2023, p. 4). However, the same formal inference principles apply to subordinate intersection null hypotheses (e.g., $H_{0,1} \cap H_{0,2}$) that have been selectively reported from an "actual" superordinate intersection null hypothesis (e.g., $H_{0,1} \cap H_{0,2} \cap H_{0,3}$). For a discussion and illustration, see Rubin (2024b, pp. 59–60).

6. Figure 1 is based on Rubin's (2025) Figure 1. Also note that a similar formal inference argument can be made to demonstrate that Type I error rates are not inflated in the case of optional stopping (Rubin, 2024b, pp. 61–62; cf. Mayo, 2018, pp. 43–44).

7. Mayo (2018, p. 14) proposed an "argument from coincidence" as a "positive counterpart" to the minimal severity requirement. As Mayo (2010) explained, "the most familiar canonical exemplar of an argument from coincidence is statistical null hypothesis testing … [Here,] the null hypothesis, $H_0$, asserts that the effect or observed agreement is 'due to coincidence'" (p. 85; see also Fisher, 1956, p. 35). When testing an intersection null hypothesis, the "coincidence" is that at least one test would yield a significant result when all constituent null hypotheses are true. If the familywise error rate is suitably low, then there is a low probability of encountering this coincidence. In this case, we can say that the union-intersection test has a high "capacity" to detect an error, and we are licensed to argue against this "coincidence" to the absence of error in the case at hand. Crucially, however, when testing an intersection null hypothesis, the case at hand is an *intersection* null hypothesis, not an *individual* hypothesis.

8. Mayo uses the term "inference" to refer to a claim that has been detached from an inferential argument (e.g., "*x* is gray" in the case of "all cats are gray, and *x* is a cat; therefore, *x* is gray"; Mayo, 2011, pp. 84–85; Mayo, 2018, p. 65). To be warranted in the error statistical approach, a statistical inference must be detached after the argument's premises have been adequately checked and affirmed as being "approximately true" (Mayo, 2011, p. 96; Mayo & Spanos, 2011, pp. 189, 193; see also Mayo, 1996, p. 435; Mayo, 2018, pp. 84, 441–442). In contrast, in the formal inference approach, a *statistical* inference cannot be detached from its inferential argument because it is subjunctively conditioned on a test's suppositional i-assumptions (Kass, 2011, pp. 5–6, 7; see also Fisher, 1956, p. 44; Hennig, 2023, p. 24). Only a nonstatistical *substantive* inference can be detached as a more general standalone conclusion. Nonetheless, following Fisher (1959, p. 21), a statistical inference can be regarded as a suppositional "as if" inductive inference. Specifically, it can be argued that *if* a test's i-assumptions *were* true, *then* a significant result *would* license the provisional inference that the study's current sample does not belong to a specified hypothetical infinite null population that contains no known relevant subsets (Fisher, 1956, pp. 32–33, 57, 109; Fisher, 1959, pp. 21–24). Despite its name, this

"inductive reasoning" is consistent with Popper's deductive approach because it is conceptualized as a "statement of uncertainty" about the observed result (Fisher, 1956, pp. 57, 108–110; Fisher, 1959, p. 22; see also Greenland, 1998, pp. 545–546) rather than an inductive prediction about future results during "repeated sampling from the same population" (Fisher, 1956, p. 77; cf. Mayo, 2018, p. 83).

9. The difference between a significant result and a nonsignificant result may not itself be significant (Gelman & Stern, 2006). Hence, at this stage, we cannot draw any conclusions about the potential moderating effect of jelly bean color.

10. In order to avoid questions about why they only tested *green* jelly beans, a researcher may deliberately omit jelly bean color from the specification of their statistical inference. In the formal inference approach, this personally biased respecification does not impact the validity of the associated Type I error rate, which is subjunctively conditioned on the researcher's formally reported inference. Like all statistical inferences, the researcher's respecified inference may include irrelevant variables and/or exclude relevant variables (Rubin, 2025, p. 7, Footnote 5). Hence, when interpreting statistical inferences, it is important to appreciate "that the population in question is hypothetical, that it could be defined in many ways, and that the first to come to mind may be quite misleading" (Fisher, 1956, p. 78). In the present case, future tests may either confirm that color is irrelevant or reveal it to be a previously hidden moderator that has caused a systematic error in the substantive inference that "eating jelly beans causes acne" (i.e., only *green* jelly beans cause acne). However, this revelation would not invalidate the researcher's original statistical inference, which requires us to accept a "postulate of ignorance" (i-assumption) about the presence of relevant subsets (e.g., *green* jelly beans) within a hypothetical infinite population (i.e., "jelly beans"; Fisher, 1959, p. 23). This postulate allows us to construe the current sample as a random sample and to attribute uncertainty about the statistical inference to random sampling error independent from systematic error (Fisher, 1956, pp. 32–33, 57, 109; Fisher, 1959, pp. 21–24). Note that if the discovery of previously hidden moderators and relevant subsets were to invalidate prior statistical inferences, then most of our prior statistical inferences would be invalid.

# References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 357–366. https://doi.org/10.1177/2515245918773742

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485–485. https://doi.org/10.1136/bmj.311.7003.485

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics, 24*, 295–300. https://doi.org/10.2307/1267823

Bishop, D. V. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568*(7753), 435–436. https://www.nature.com/articles/d41586-019-01307-2

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2*(2), 161–172. https://doi.org/10.1037/1082-989X.2.2.161

Firestein, S. (2016, February 14). Why failure to replicate findings can actually be good for science. *L A Times*. https://www.latimes.com/opinion/op-ed/la-oe-0214-firestein-science-replication-failure-20160214-story.html

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

Fisher, R. A. (1959). Mathematical probability in the natural sciences. *Technometrics, 1*(1), 21–29. https://doi.org/10.1080/00401706.1959.10489846

García-Pérez, M. A. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology*, 8, 100120. https://doi.org/10.1016/j.metip.2023.100120

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Goeman, J. (2022, May 12). Review of: When to adjust alpha during multiple testing. *Qeios*. https://doi.org/10.32388/48X0KD

Greenland, S. (1998). Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology, 27*(4), 543–548. https://doi.org/10.1093/ije/27.4.543

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, 186(6), 639−645. https://doi.org/10.1093/aje/kwx259

Hennig, C. (2010). Mathematical models and reality: A constructivist perspective. *Foundations of Science, 15,* 29–48. https://doi.org/10.1007/s10699-009-9167-x

Hennig, C. (2023). Probability models in statistical data analysis: Uses, interpretations, frequentism-as-model. In B. Sriraman (Ed.), *Handbook of the History and Philosophy of Mathematical Practice* (pp. 1–49)*.* Springer. https://doi.org/10.1007/978-3-030-19071-2_105-1 [Accessed as preprint version: https://arxiv.org/abs/2007.05748]

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. Wiley. http://doi.org/10.1002/9780470316672

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science, 26*(1), 1. https://doi.org/10.1214/10-STS337

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3)*,* 196–217. http://doi.org/10.1207/s15327957pspr0203_4

Lakatos, I. (1978). *The methodology of scientific research programmes (Philosophical Papers, Volume I)*. Cambridge University Press.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review, 62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. (2008). How to discount double-counting when it counts: Some clarifications. *The British Journal for the Philosophy of Science, 59*(4), 857–879. https://doi.org/10.1093/bjps/axn034

Mayo, D. G. (2010). Can scientific theories be warranted with severity? Exchanges with Alan Chalmers. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 73–112). Cambridge University Press.

Mayo, D. G. (2011). Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)? *RMM, 2,* 79–102. http://hdl.handle.net/10919/80466

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D. G. (2025). Severe testing: Error statistics versus Bayes factor tests. *British Journal for the Philosophy of Science.* https://doi.org/10.1086/736950

Mayo, D. G., & Cox, D. (2010). Frequentist statistics as a theory of inductive inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 247–275). Cambridge University Press.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323–357. https://doi.org/10.1093/bjps/axl003

Mayo, D. G., & Spanos, A. (2011). Error statistics. In S. Bandyopadhyay, & M. R. Forster (Eds.), *Handbook of philosophy of science: Philosophy of statistics* (Vol. 7, pp. 152–198). Elsevier.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965p li0102_1

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Lawrence Erlbaum.

Miller, J. D., Phillips, N. L., & Lynam, D. R. (2025). Questionable research practices violate the American Psychological Association's Code of Ethics. *Journal of Psychopathology and Clinical Science, 134*(2), 113–114. https://doi.org/10.1037/abn0000974

Munroe, R. (2011). *Significant*. https://xkcd.com/882/

Murphy, S. L., Merz, R., Reimann, L. E., & Fernández, A. (2025). Nonsignificance misinterpreted as an effect's absence in psychology: Prevalence and temporal analyses. *Royal Society Open Science,* 12242167. http://doi.org/10.1098/rsos.242167

Nagy, T., Hergert, J., Elsherif, M. M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., Wang, Y. A., Scharfenberg, D., Tyson, G., Yang, Y.-F., Stvortsova, A., Alarie, S., Graves, K., Sotola, L. K., Moreau, D., & Rubínová, E. (2025). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science, 8*(3). https://doi.org/10.1177/25152459251348431

Neyman, J. (1950). *First course in probability and statistics.* Henry Holt.

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability (2ⁿᵈ ed.).* Graduate School, US Department of Agriculture. https://doi.org/10.22004/ag.econ.327287

Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics, 8,* 13–46. https://doi.org/10.1002/cpa.3160080103

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika 20A*(1/2)*,* 175–240. http://doi.org/10.2307/2331945

Parker, R. A., & Weir, C. J. (2020). Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification. *Clinical Trials, 17*(5), 562–566. https://doi.org/10.1177/1740774520941419

Pickett, J. T., & Roche, S. P. (2018). Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Science and Engineering Ethics*, *24*, 151–171. https://doi.org/10.1007/s11948-017-9886-2

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*(1), 159–163. https://doi.org/10.1037/0033-2909.102.1.159

Popper, K. R. (1966). *The open society and its enemies* (5th ed.). Routledge.

Popper, K. R. (1983). *Realism and the aim of science: From the postscript to the logic of scientific discovery.* Routledge.

Popper, K. R. (1994). *The myth of the framework: In defence of science and rationality.* Psychology Press.

Popper, K. R. (2002). *The logic of scientific discovery.* Routledge.

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge.* Phoenix Books.

Reis, D., & Friese, M. (2022). The myriad forms of *p*-hacking. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 101–121). Springer. https://doi.org/10.1007/978-3-031-04968-2_5

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics, 24*(2), 220–238. https://doi.org/10.1214/aoms/1177729029

Rubin, M. (2017a). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*(3), 269–275. https://doi.org/10.1037/gpr0000123

Rubin, M. (2017b). 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*(4), 321–329. https://doi.org/10.1037/gpr0000135

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology, 16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376

Rubin, M. (2021a). There's no need to lower the significance threshold when conducting single tests of multiple individual hypotheses. *Academia Letters,* 610. https://doi.org/10.20935/AL610

Rubin, M. (2021b). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese, 199*, 10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Rubin, M. (2022). The costs of HARKing. *British Journal for the Philosophy of Science, 73*(2), 535–560. https://doi.org/10.1093/bjps/axz050

Rubin, M. (2024a). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology, 10,* 100140. https://doi.org/10.1016/j.metip.2024.100140

Rubin, M. (2024b). Type I error rates are not usually inflated. *Journal of Trial and Error, 4*(2), 46–71. https://doi.org/10.36850/4d35-44bd

Rubin, M. (2025). Preregistration does not improve the transparent evaluation of severity in Popper's philosophy of science or when deviations are allowed. *Synthese, 206,* 111. https://doi.org/10.1007/s11229-025-05191-4

Rubin, M., & Donkin, C. (2024). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology, 37*(8), 2019–2047. https://doi.org/10.1080/09515089.2022.2113771

Selvin, H. C. (1958). Durkheim's suicide and problems of empirical research. *American Journal of Sociology, 63(*6), 607–619. https://doi.org/10.1086/222356

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Spanos, A. (2010). Theory testing in economics and the error-statistical perspective. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 202−246). Cambridge University Press.

Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese, 192*(11), 3533–3555. https://doi.org/10.1007/s11229-015-0744-y

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of *p*-hacking strategies. *Royal Society Open Science, 10*(2), 220346. https://doi.org/10.1098/rsos.220346

Waller, J. (2018). Division. In R. Arp, S. Barbone & M. Bruce (Eds.), *Bad arguments* (pp. 259–260). Wiley. https://doi.org/10.1002/9781119165811.ch56

Weisberg, H. I. (2014). *Willful ignorance: The mismeasure of uncertainty.* Wiley.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005