Type M Error: Consequences of Selection Bias when the Null is False

Nathan Mitchell

Abstract

One of the fundamental consequences of null hypothesis significance testing is the idea of a false positive. Many practices, such as p-hacking and selective reporting, that researchers or journal editors enact increase the rate at which the results being published may be classified as such. It has been argued, however, that the null hypotheses being tested against do not make practical sense as they are always false. In an attempt to move away from strict hypothesis testing, effect sizes can be estimated. The effects of p-hacking and selective reporting do not go away if we stop considering false positives but simply take another form — effect size bias.

Classical inference is based on the idea that the sample we have taken was collected from a larger population and that it was one of many possible samples that could have been taken. Though this is largely a thought experiment, it gives us the theoretical basis for many processes carried out under classical framework. One such process is statistical inference, which can be thought of as a way to go beyond the sample we have collected and make claims about the population as a whole. Within the idea of statistical inference comes the idea of null hypothesis significance testing. Hypotheses are organized by how they characterize the expected difference between the possible study groups [1].

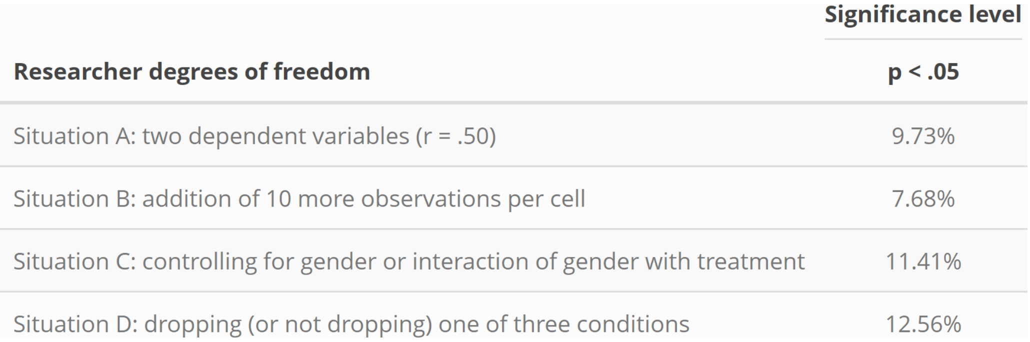
To start, we can look at the null and alternative hypotheses. The nil-null hypothesis states that there is no association between the predictor(s) and the outcome at the population level. The alternative hypothesis, then, proposes that there is an association between these variables [1]. Derived from this are one- and two-tailed alternative hypotheses. A one-tailed, or one-sided, alternative hypothesis allows us to specify a direction of interest, either positive or negative. A two-sided alternative hypothesis does not state a direction of interest and thus only cares if a difference exists [1].

We don't directly test against the alternative hypothesis, however. If you reject the null you favor the alternative. The decision that is made is whether or not we should reject the null hypothesis. This rejection is done at some level, α, typically 0.05, by generating a test statistic. This test statistic can then be used to calculate a p-value, which is the probability of calculating a test statistic at least as large as the one we calculated if the null was true and a new sample was tested. In typical practice, if this value is less than or equal to 0.05, we reject the null hypothesis. If it's greater than 0.05, we fail to reject our null hypothesis. This rejection tells us that we have enough evidence to claim that, if the null were true, the results shown are unlikely to have occurred by chance.

While performing these tests, however, there is always the possibility of making the wrong decision. One way this can be done is known as a Type I error, or a false positive, and might occur when we have rejected the null hypothesis. Rejecting the null hypothesis would be the incorrect decision if the null were true at the population level. If all statistical procedures were followed correctly, the false positive rate should be 0.05, our α level. Depending on the α level chosen, this rate may increase or decrease. Under a classical lens, this means that, if the null is true, under repeated sampling, our method should fail to reject the null 95% of the time — hence it only makes the wrong decision 5% of the time. The other type of error, the Type II error, or false negative, might occur when we fail to reject the null hypothesis but, at the population level, the null is false.

As mentioned above, a false positive rate of 0.05 can be achieved if standard protocol is followed. Researchers, however, do not always perform everything correctly. One example of this is known as p-hacking, or "researcher degrees of freedom" [3]. A researcher may p-hack by performing multiple different statistical tests on the same set of data — or even different subsets of the data, often derived by dropping outliers — selectively choosing which response variable to use, or simply getting rid of unwanted data until a significant result is found; all after an initial analysis has been conducted. It is imperative for these decisions to be made before the data have been analyzed as the interpretation of the p-value infers this is the case [2].

It should be noted, however, that these practices may have been performed implicitly; the researcher unaware an error had been made at all due to ambiguity when collecting and analyzing data such as whether or not more data should be collected or what to do with outliers. Such questions might not have been possible to ask during the planning process. However implicit or justified, these practices have been shown to lead to results of questionable validity. When researchers only report what led to them obtaining statistical significance, even though other methods were attempted, the likelihood that at least one of these tests produced a falsely positive result rises from the desired level [3].



	Significance level
Researcher degrees of freedom	p < .05
Situation A: two dependent variables (r = .50)	9.73%
Situation B: addition of 10 more observations per cell	7.68%
Situation C: controlling for gender or interaction of gender with treatment	11.41%
Situation D: dropping (or not dropping) one of three conditions	12.56%

Table 1: The likelihood of obtaining a falsely-positive result given different forms of research design flexibility.

To test this, Simmons et al. (2011) ran simulations that sought to model some practices that researchers may not know are affecting their results. The four practices that they wanted to model were "...(a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions"[3]. I was able to recreate their simulations and have displayed the results of this recreation in Table I.

This table shows the percentage of results that obtained statistical significance, $p < 0.05$, in simulations that each utilized 250,000 simulated samples. All observations were independently drawn from a normal distribution with a variance of one. The baseline is a test between two conditions, each with 20 observations. Here, we are working under the assumption that the null is true, thus any significant result found is classified as a false positive. According to our α level, 5% of the simulated results are expected to be false positives.

Situation A represents t-tests that were performed between two dependent variables that are set to be correlated at $r = 0.50$. The first and second are on the dependent variables and the third is on the average of the two. The smallest p-value of the three is kept. Situation B represents a researcher conducting a t-test between two dependent variables. If the researcher finds significance, this p-value is kept. If not, 10 more observations per condition are collected and another test is performed. The p-value from this test is kept regardless of its significance status.

The results from Situation C were found by performing a t-test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction. Here, "gender" is a variable that has a 0.50 probability of being assigned a 1 (or male). The smallest p-value from these three tests is kept. The first two are from testing if the condition is significant, and the third is from testing if the interaction is significant.

Situation D generates three equal conditions of size 20 and tests them against each other to find the first three p-values. Then, a linear regression model is fit on all three predictors against low = -1, medium = 0, and high = 1. The smallest of the four p-values is kept.

The takeaway from these simulations is that when p-hacking occurs, the rate at which false positives are observed can increase dramatically from the desired level of 5%.

Building on the idea of Situation B, Simmons et al. (2011) ran another simulation. The basis of this simulation was questioning what would happen to false positive rates if the data collection process ended when significance was reached. To simulate this, each of two conditions was given a starting sample size of either 10 or 20. Each observation was generated from a normal distribution with a mean of zero and a variance of one. A t-test was performed between the two conditions and the resulting p-value was pulled. If significance was found, data collection stopped, and the significant p-value was recorded. If p was greater than 0.05, either one, five, 10, or 20 more samples were collected per condition and another t-test was performed. This process continued either until significance was found or there were 50 total observations per condition. Figure I shows the results of a recreation of this simulation. The baseline percent of false positives that should be seen is shown by the dotted line.

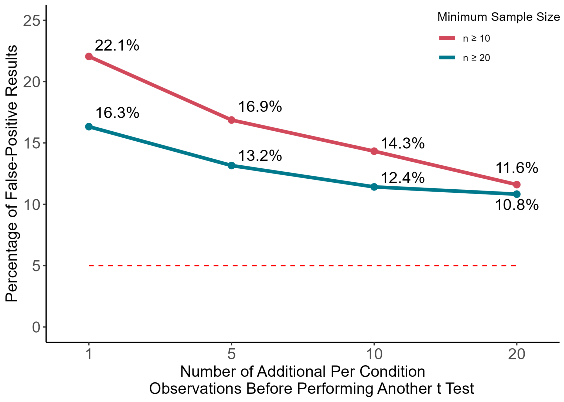


Figure I: The Likelihood of obtaining a falsely-positive result when data collection ends when a significant result is reached ($p < 0.05$, represented by the dashed line).

As can be seen, as the number of times the researcher tests for significance decreases, and the number of additional per-condition observations collected increases, so does the false positive rate. When the hypothetical researcher started with a sample size of 10 per condition and tested for significance after every new additional sample collected per condition, the false positive rate shot up to just over 20%, which is well above the 5% false positive rate that is expected. This simulation was originally performed because Simmons et al. (2011) found that roughly 70% of researchers admitted to having stopped data collected based on the results of an analysis that was performed before the data collection process had been completed. They also noted that many researchers believed that this process did not greatly increase the false positive rate. As can be seen, that does not appear to be the case. Thus, these results represent realistic false-positive estimates of published research.

A natural conclusion of increased false positive rates is that many published research findings are false. This effect is exaggerated when we consider the act of selective reporting which occurs when studies that do not show significant results are withheld from publication. This can be done by journals and their editors or by individual researchers themselves [6]. If journals are looking to publish results that meet the $p < 0.05$ threshold, researchers may have an incentive to find positive results. Thus, if journals require significant results and researchers are *trying* to get them, it can easily be seen how many published results are the result of false positives. This idea has been explored before, notably by Ioannidis (2005), where he used bias, statistical power, and the pre-study odds of a finding being true to explain this phenomenon. For reference, the statistical power of a study is the probability of correctly rejecting a false null hypothesis. To corroborate the claim that most published research findings are false, Ioannidis cites the increasing rate at which newly found evidence contradicts old results and the decreasing rate at which studies are able to be replicated.

The idea of false positives and false research findings relies on one basic assumption — an assumption that is built into the process of null hypothesis significance testing and is thus often overlooked. To see this, recall that a sample is meant to mirror the population that it comes from. Because of random sampling error, any measurement taken from this sample will differ from its population-level counterpart by some amount [7]. The same logic can be applied when comparing two samples to one another. The assumption that is made when conducting a hypothesis test, an assumption that is always violated, is that it is possible for the null to be true. Formal language forbids us to ever actually *say* that the null is true or that we *accept* the null being true, but by testing against it in the first place we allow ourselves to believe that it *might* be true or there would be no reason to perform the test.

This idea, though overlooked by some, has not been overlooked by all. As noted by Cohen (1990), "The null hypothesis, taken literally,... is *always* false in the real world." If we are willing to believe this, we run into a problem with our Type I and Type II error rates. If the null is always false, if we fail to reject the null we know we have made the wrong decision. Thus, every time that we fail to reject we must be committing a Type II error. Type I errors, on the other hand, could never possibly occur as, by definition, they require the null to possibly be true. Suddenly, our aforementioned Type I error rate of 5% refers to something that cannot happen.

When it is no longer possible to commit a Type I error, the results that Ioannidis (2005) and Simmons et al. (2011) displayed — how increased false positive rates due to p-hacking/selective reporting — become impertinent. Even though we are no longer quantifying a mistake as a false positive, the practices that they examine still must have an effect on published results. To see this effect, we can examine a different, commonly-used statistical method — effect size estimation; an alternative to hypothesis testing that answers the same question as hypothesis testing but provides more, useful information. When conducting a study, whether or not a treatment has *any* effect is not typically of utmost interest. The *size* of this effect, however, is [7]. Thus, the estimated size of an effect is of direct interest.

An effect size is a quantitative measure of the magnitude of an effect such as the difference between two groups or how strongly two variables are correlated (Button et al., 2013). It follows that an estimation of this effect attempts to determine what this value takes on at the population level. This estimate is accompanied by a confidence interval which tells how "certain" the estimate is [7]. For instance, if the confidence interval is wide around the estimate, there is less certainty behind the estimate than if the interval was narrow. All values present within the confidence interval are values for which the population-level estimate *could* take on — that is, they are values we would fail to reject. There are many different effect sizes that are often estimated and used but the one that will be used here is Cohen's d, the standardized mean difference. To calculate this effect size, the difference between two means is calculated and then divided by the standard deviation. Formally, this is written as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Now that we have switched to estimating effect sizes using Cohen's d, we can re-examine p-hacking and selective reporting. The former stays the same, but selective reporting takes a slightly different form. Large, positive effect size estimates — those that are considered "clinically important," and are often statistically significant — are published in favor of results that have a small magnitude or are in a negative direction [6]. As a result, the effect sizes that are getting published — the ones that are actually *seen* —may have been exaggerated and no longer reflect their population-level counterparts.

The extent to which these estimates are exaggerated is a product of different factors that work in unison: sample size and the size of the effect being studied. To illustrate this, consider a study that utilizes a small sample size, n = 10, and aims to study a small effect, $D = 0.1$. This hypothetical study would have very low statistical power — roughly 5.51% — because of the small effect size being estimated and the small sample size. Low power means that the probability of the effect size calculated *accurately reflects* the size of the effect — assuming that there is an effect to begin with — decreases [8]. Whether or not this effect size is considered to be statistically significant can roughly be determined by considering if $|d/s| > 2$, where s is the standard error of the estimated effect size [9].

Assume that the effect size was found to be significant and was published. If this study was trying to estimate $D = 0.1$, their estimate, d, *should* be close to 0.1. Thus, in order for them to have reached statistical significance, they would have needed to have a small s to force $|d/s|$ to be greater than 2. Because this study used such a small sample size, however, we know that the standard error of their estimate must have been large — too large to garner statistical significance. Thus, it must have been the case that d was exaggerated. The phenomenon in which researchers find statistical significance but end up with an exaggerated effect size has been called the "winner's curse" [11].

There are a few variations on this that should be explored further. First, we can examine, on an effect that is fixed, what happens when sample size increases. With an increase in sample size comes a decrease in the standard error of the estimate. Consequently, when sample sizes are small, estimates will need to be large to counteract their standard errors, but when sample sizes are large, estimates no longer need to counteract their standard errors to become significant and will thus be less exaggerated. Next, we can examine what happens when we fix sample size and increase the size of the effect being studied. In this scenario, as the size of the effect increases, it becomes easier to reach significance because the effects being estimated are getting larger relative to their standard errors. Thus, small population-level effects will become exaggerated in order to reach significance but larger effects will not need that exaggeration to make them significant.

Knowing this, it can be said that the extent to which an estimated effect size is exaggerated is a function of the statistical power of the study that estimated that effect. As power — and consequently sample sizes and population-level effect sizes — increases, estimates become less exaggerated. Similarly, as power decreases, estimates become more exaggerated.

This exaggeration can be measured by calculating the Type M error, a measurement thought up by Andrew Gelman and John Carlin [9]. For a given effect size, say Cohen's d, the Type M error can be calculated as follows:

$$\frac{|d\ given\ p{<}0.05|}{d_{hypothesized}}$$

This represents how much a result has been exaggerated from its hypothesized value, not the expected value of the estimation. For this calculation, we only consider the effect size estimates that have been deemed statistically significant. One simple reason for this is that the published research findings, the ones that have reached statistical significance, are typically the ones that we get to know about in practice.

In order to simulate how exaggerated effect size estimates might be, it is convenient to believe that the nil null is always false. Mathematically speaking, if there is no effect at the population level, $d = 0$, there is no way to divide by the hypothesized effect size. There is, however, statistical theory to give credit to this assumption. John Tukey started "The Philosophy of Multiple Comparisons" [10] off with "Statisticians classically asked the wrong question — and were willing to answer with a lie, one that was often a downright lie. They asked 'Are the effects of A and B different?' and they were willing to answer 'no'." Just as Cohen claimed this for hypothesis tests, Tukey claimed that, in practice, two effects will never truly be the same. Tukey backs up his claim by noting that two effects will always be different at some level due to an increase in measuring technology. Surely if two effects are not different at one decimal place, a more precise measurement can be taken to eventually find a difference.

As discussed so far, exaggeration of effect size estimates exists when power is low and this exaggeration decreases as power increases. Figure II displays the results of a simulation that demonstrates this phenomenon. The baseline condition is a two-sample t-test with each condition generated from a normal distribution; each receiving an equal sample size — one of 10, 20, 30, 50, 100, or 200 — and a variance of one. The first group would receive a mean of either 0.1, 0.2, 0.5, or 0.8, and the second a mean of — thus inducing a population-level difference in effect sizes of one of those measures — i.e. the nil null is false. From there, a t-test is conducted. Regardless of the significance status, the absolute value of the effect size is kept. Once the data has been collected, the mean of the effect sizes is taken and divided by the true, population-level effect size. The dashed line represents an exaggeration of 1, i.e. the results were not exaggerated and reflect their population-level counterpart.

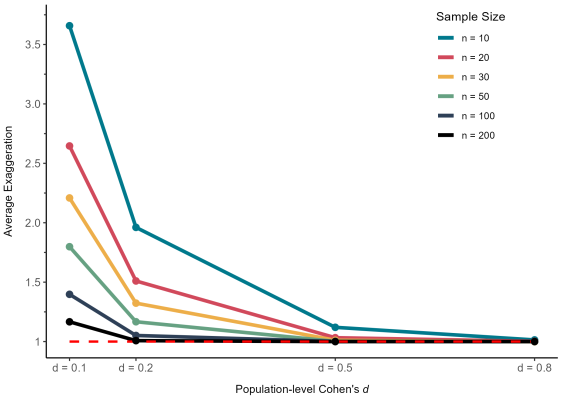


Figure II: Average effect size exaggeration for increasing sample sizes and Cohen's d — given the null is false.

As can be seen, and as has been described, when power is low — when a study has a small sample size and is studying a small effect — effect sizes are exaggerated. This is due to the random sampling bias present in the observations when a small sample size is used. As power increases, this exaggeration decreases and effect sizes are being estimated with little to no error, on average. It should be noted that Figure II does not represent the Type M error as results that were not considered statistically significant were used in the calculation of the average exaggeration. Figure II serves as a baseline for expected Type M error, which can be seen in Figure III below.

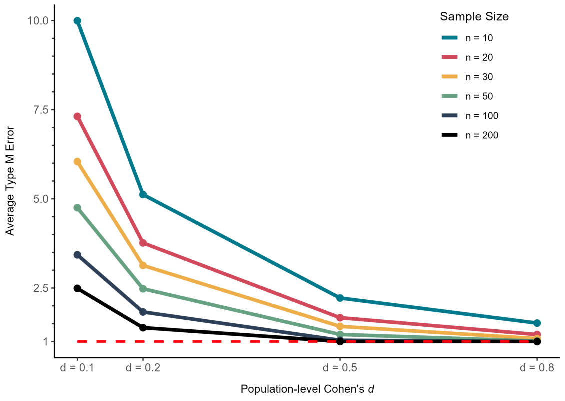


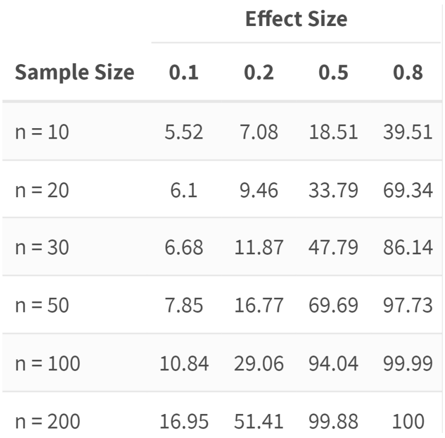
Figure III: Average Type M error for increasing sample sizes and true Cohen's *d* given the null is false and significance has been selected for.

This figure represents how exaggerated results can be in the presence of selective reporting. Once again, the dashed line represents the baseline of no exaggeration. The same process was carried out in Figure III as in Figure II — the only difference being that only the statistically significant results were kept. This, then, is a measure of the average expected Type M error as it is written.

As can be seen, selection for significance dramatically increases the factor by which estimates are exaggerated. Previously, studies with a sample size of 10 per condition that aimed to estimate an effect of 0.1 were exaggerated, on average, by a factor of 3.6. When only the studies that reached publication are considered, however, that exaggeration factor rises to 10. Instead of finding an effect of 0.1, studies, on average, were estimating that this effect was of size 1.

Thus, even when the null is false, and a true, population-level effect exists, the estimates that are actually seen may be exaggerated from the hypothesizes effect of interest.

These results follow from what Ioannidis (2005) conjectured in two of his corollaries that followed from his claim that most published research findings are false. His first corollary states that “The smaller the studies conducted in a scientific field, the less likely the research findings to be true,” and his second states that “The smaller the effect sizes in a scientific field, the less likely the research findings to be true” [5]. When combined, we get that studies with small sample sizes and studies that aim to measure small effects have a low chance of finding the truth. His research was concerned about false positive rates, but it can be seen that those same conditions lead to exaggerated results, as was shown in Figures II and III.



	Effect Size			
Sample Size	0.1	0.2	0.5	0.8
n = 10	5.52	7.08	18.51	39.51
n = 20	6.1	9.46	33.79	69.34
n = 30	6.68	11.87	47.79	86.14
n = 50	7.85	16.77	69.69	97.73
n = 100	10.84	29.06	94.04	99.99
n = 200	16.95	51.41	99.88	100

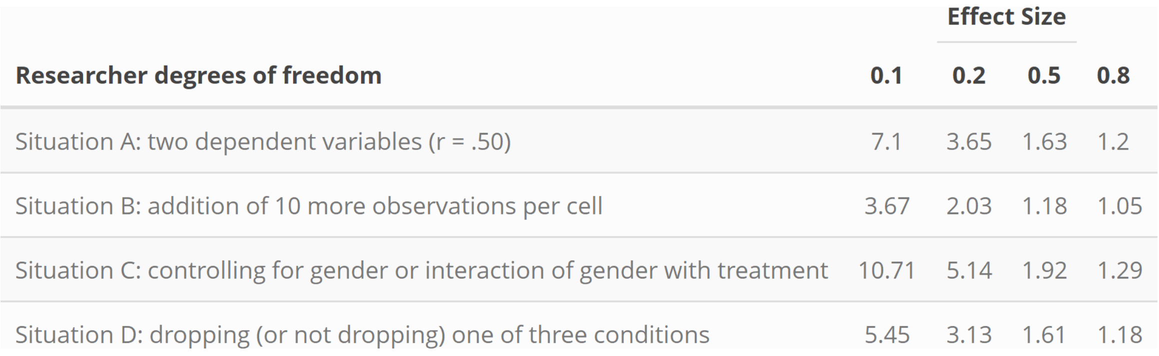
Table 2: Statistical Power - Figure III

To corroborate the claim that power is the driving force behind the Type M error, we can examine the results shown in Table 2. The results shown in this table represent a power analysis that was done for each trial in Figure III. As can be seen, as the sample size increases, and as the size of the effect being studied gets larger, the estimated power grows. When comparing these results to the data shown in Figure III, the pattern becomes clear. As power increases, the Type M error decreases.

If the results from one study that was concerned with false positives can be replicated in terms of effect size estimation exaggeration, the same treatment can be applied to others. Namely, the two simulations from Simmons et al. (2011) on researcher degrees of freedom and flexibility in sample size — the two simulations that were recreated above and shown in Table 1 and Figure I. Now, instead of calculating the false positive rate when the null is true, as has been posed in many studies, we can examine the amount by which effect size estimates are exaggerated when the null is false.

To start, we can look at what happens when researchers exert their degrees of freedom, or *p*-hack, during the data collection and analysis processes. Table 3 displays a recreation of the results found in Table I applied to effect size estimation. Each of the four situations was conducted in the same manner but, instead of just pulling a *p*-value, an estimate of Cohen's *d* was taken and the average Type M error was calculated.

Similarly to how false positive rates were inflated when researchers *p*-hacked, it can be seen that *p*-hacking inflates the estimated size of an effect for any given effect size. It can also be seen that the same rules for power-based exaggeration apply here. Here, however, the only variable changing is the population-level effect size. Each situation sees the highest Type M error when the population-level effect size is at its smallest — 0.1 — and this exaggeration decreases as the size of the effect increases. Something interesting to note is that what caused the highest level of false positives in the original Situation D did not correlate to the highest Type M error seen.



	Effect Size			
Researcher degrees of freedom	**0.1**	**0.2**	**0.5**	**0.8**
Situation A: two dependent variables (r = .50)	7.1	3.65	1.63	1.2
Situation B: addition of 10 more observations per cell	3.67	2.03	1.18	1.05
Situation C: controlling for gender or interaction of gender with treatment	10.71	5.14	1.92	1.29
Situation D: dropping (or not dropping) one of three conditions	5.45	3.13	1.61	1.18

Table 3: Estimated Type M error given different forms of research design flexibility.

Simmons et al. (2011) also simulated how increased flexibility in data collection led to increased false positive rates. The results from a recreation of their simulation can be seen in Figure I. To recreate this in terms of the Type M error, instead of keeping track of the p-values, a measure of Cohen's d was taken and recorded if statistically significant. The same process of data flexibility was explored but for $d = 0.1$ only. The results of this recreation can be found in Figure IV.

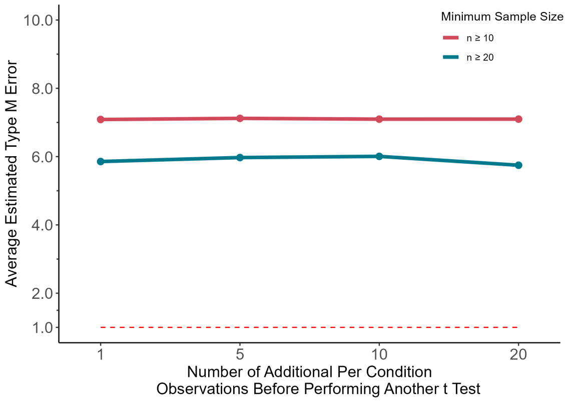
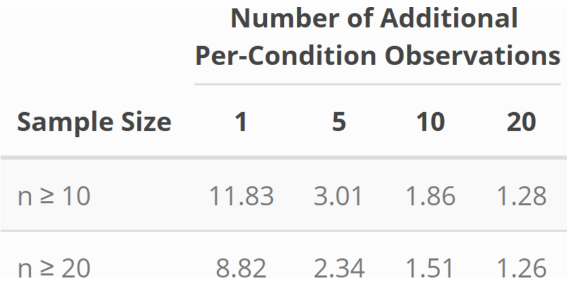


Figure IV: Average estimated Type M error for different per-condition sample sizes and additional per-condition observations before performing another t-test — given the null is false.

Interestingly, the results from this recreation tell a different story than the original simulation. Previously, as the number of additional per-condition observations increased, the percentage of false positives decreased for both a starting per-condition sample size of 10 and 20. When effect size estimates are considered, it does not appear that the number of additional per-condition observations affected how exaggerated the estimates were.

Despite this, there is still clear evidence that flexibility in sample size has a large effect on how exaggerated effect size estimates are. Though the Type M error stays constant, the value it is staying constant at for a per-condition sample size of 10 is roughly 7.1, meaning that, on average, effect size estimates are just over seven times as large as they should be. Of course, as the starting per-condition sample size increases, this exaggeration factor goes down. It is also expected that, as the population-level Cohen's d increases, the effects of flexibility in sample size will decrease as has been shown previously.

To accompany Figure IV is Table 4, shown below. This table shows the average number of attempts that were needed to reach significance in each trial. As can be seen, as the population-level Cohen's d increases and as the minimum sample size increases, the number of attempts that are needed to reach significance decreases. The first two rows of Table 2 show the power of each of these hypothetical studies.



	Number of Additional Per-Condition Observations			
Sample Size	**1**	**5**	**10**	**20**
n ≥ 10	11.83	3.01	1.86	1.28
n ≥ 20	8.82	2.34	1.51	1.26

Table 4: Average Number of Attempts Needed to Reach Significance - Figure IV

An additional corollary that Ioannidis (2005) mentioned is relevant to both of these scenarios — researcher degrees of freedom and flexibility in sample size. His fourth corollary states that "The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true." This result was demonstrated exceptionally by Simmons et al. (2011) and can be further generalized to effect size estimation and the Type M error similarly to his first and second corollaries. The results from Table 3 and Figure V show this result. As researchers exert flexibility in their designs and data-collection methods, the amount by which effect size estimates are exaggerated increases.

From the results discussed, it follows that the Type I and Type II errors, while useful in examining why research is often unable to be replicated, do not represent the full picture when it comes to analyzing the consequences of p-hacking. When researchers are not honest about their data collection and analysis practices, they not only run the risk of generating results that are false positives but also increase how exaggerated their estimates are from the effect attempting to be estimated. Even when the effect being tested for *exists* — the nil null is false — the estimate produced may be exaggerated to a substantial degree. The same result can be seen when results are selectively reported. The natural conclusion that follows from these results is that statistical power plays a large role — a larger role than many may realize — in the estimates a study produces.

REFERENCES:

[1] Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. Industrial psychiatry journal, 18(2), 127–131. https://doi.org/10.4103/0972-6748.62274

[2] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS biology, 13(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

[3] 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

[4] Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

[5] Ioannidis J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

[6] Nair A. S. (2019). Publication bias - Importance of studies with negative results!. Indian journal of anaesthesia, 63(6), 505–507. https://doi.org/10.4103/ija.IJA_142_19

[7] Borenstein M. (1997). Hypothesis testing and effect size estimation in clinical trials. Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology, 78(1), 5–16. https://doi.org/10.1016/S1081-1206(10)63363-7

[8] Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. Royal Society open science, 4(2), 160254. https://doi.org/10.1098/rsos.160254

[9] Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641-651. https://doi.org/10.1177/1745691614551642

[10] Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. Statistical Science, 6(1), 100–116. http://www.jstor.org/stable/2245714

[11] Button, K., Ioannidis, J., Mokrysz, C. et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14, 365–376 (2013). https://doi.org/10.1038/nrn3475