Rethinking Type S and M Errors

Daniel Lakens[1], Cristian Mesquida[1], Gabriela Xavier-Quintais[2], Sajedeh Rasti[1], Enrico Toffalini[3], and Gianmarco Altoè[4]

[1]Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

[2]Faculty of Sciences, University of Lisbon

[3]Department of General Psychology, University of Padova

[4]Department of Developmental Psychology and Socialisation, University of Padova

Author Note

Daniel Lakens https://orcid.org/0000-0002-8393-5316

Cristian Mesquida https://orcid.org/0000-0002-1542-8355

Gabriela Xavier-Quintais https://orcid.org/0000-0003-4896-1225

Sajedeh Rasti https://orcid.org/0009-0007-3416-7692

Enrico Toffalini https://orcid.org/0000-0002-1404-5133

Gianmarco Altoè https://orcid.org/0000-0003-1154-9528

A reproducible version of this manuscript is available at https://github.com/Lakens/rethinking_type_s_and_m_errors. The authors have no conflict of interest to declare.

Correspondence concerning this article should be addressed to Daniel Lakens, Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Den Dolech 2, 5600 MB, Eindhoven, Eindhoven, Email: D.Lakens@tue.nl

Abstract

Gelman and Carlin (2014) introduced Type S (sign) and Type M (magnitude) errors to highlight the possibility that statistically significant results in published articles are misleading. While these concepts have been proposed to be useful both when designing a study (prospective) and when evaluating results (retroactive), we argue that these statistics do not facilitate the proper design of studies, nor the meaningful interpretation of results. Type S errors are a response to the criticism of testing against a point null of exactly zero in contexts where true zero effects are implausible. Testing against a minimum-effect, while controlling the Type 1 error rate, provides a more coherent and practically useful alternative. Type M errors warn against effect size inflation after selectively reporting significant results, but we argue that statistical indices such as the critical effect size or bias adjusted effect size are preferable approaches. We do believe that Type S and M errors can be valuable in statistics education where the principles of error control are explained, and in the discussion section of studies that fail to follow good research practices. Overall, we argue their use-cases are more limited than is currently recognized, and alternative solutions deserve greater attention.

Keywords: Type M error, Type S error, Error Control, Hypothesis Testing, Bias Correction

Rethinking Type S and M Errors

Neyman-Pearson hypothesis testing is a widespread approach to statistical inferences in the social sciences. As a consequence of its popularity, it is commonly criticized, and critics regularly propose alternative statistical procedures to interpret results. One criticism of null hypothesis tests is that the null hypothesis is never true (Cohen, 1994). Gelman and Tuerlinckx (2000) argue that because it seems unlikely that effect sizes are ever exactly zero for continuous data, it is uninteresting to control the probability of a Type 1 error. After all, if the null is never true, researchers do not need to worry about incorrectly claiming there is an effect, when there is no effect. They propose to compute the Type S error, or 'sign error', which quantifies the long-run frequency that a significant effect in the opposite direction of the true effect size is observed. They argue that Type S error rates are "the relevant error rate for statistical analyses in the social and behavioral sciences" (p. 388). Gelman and Tuerlinckx (2000) also propose to report the Type M error, which quantifies the absolute average effect size inflation for a study design if only statistically significant results are reported. Gelman and Carlin (2014) suggest that Type M errors are useful to understand that significant effect sizes from underpowered studies are almost certain to be a huge overestimate of the true effect. Gelman and Carlin (2014, p. 644) say that problems with Type S errors become a concern "when power is less than 0.1" and Type M errors become a concern "when power is less than 0.5".

Before criticizing the proposed use of Type S and M errors, it is important to acknowledge several points of agreement. First, we agree with Gelman and Tuerlinckx (2000) that it is uninteresting to test against a null hypothesis of no effect when it is highly unlikely that the true effect size is zero. However, we believe this problem should be addressed by performing a minimal effect test for superiority against a range of values that is considered practically equivalent to zero. Such tests against a non-nill null hypothesis

have been suggested for more than half a century Nunnally (1960). We also agree with Gelman and Carlin that performing an a priori power analysis based on estimates of a previous study or an expected effect size is not best practice, and leads to bias (Albers & Lakens, 2018). However, we see no problem with what is currently considered best practice - performing an a-priori based on a smallest effect size of interest (SESOI) - just because the true effect size might be smaller than the SESOI. Trivially small effect sizes are scientifically uninteresting, and researchers should not be interested in detecting effects that are too small to matter (Lakens, 2022).

Third, we agree that publication bias is a systemic problem in the social sciences, which leads to inflated effect sizes in the scientific literature. In situations where no assumed plausible effect size under the alternative hypothesis is available, reporting the critical effect size (e.g., Perugini et al., 2025) may offer a more informative perspective than computing a Type M error, which relies on such an assumption. Finally, we recognize that it is valuable to retrospectively evaluate how informative studies in a literature are, especially in fields where good research practices are uncommon. However, we believe a sensitivity power analysis combined with bias correction methods such as p uniform (Aert & Assen, 2018) or a maximum likelihood estimate for a truncated distribution (Anderson & Maxwell, 2017; Taylor & Muller, 1996) are more insightful. Finally, the last two authors have co-authored scientific articles promoting Type S and M errors (Altoè et al., 2020; Bertoldo et al., 2022), and co-created the R package PRDA that computes Type S and M errors for different designs (Callegher et al., 2021). They were invited to co-author this article to guarantee fair and nuanced arguments by incorporating internal criticism into this project (Lakens, 2020). The last two authors found the discussions that shaped the manuscript intellectually rewarding, and the discussions confirmed the view—also expressed in the paper—that Type S and M errors, though not without limitations, can serve as useful

concepts to foster critical reflection and statistical awareness. Like many tools in scientific practice, their usefulness does not lie solely in their technical properties, but in researchers' ability to understand their logic, reflect on their limitations, and apply them thoughtfully.

Type 1, Type 2, Type 3, and Type S errors

Type 1 errors occur when the null hypothesis is true, but a statistically significant result is observed. The null hypothesis in a two-sided test is that the population effect size is 0. The alternative hypothesis is that there is an effect - either in the positive or negative direction. After observing a statistically significant result the correct claim is that the null hypothesis can be rejected, and we can act - with an error rate of α - that there is a non-zero effect. Because no directional claims are made in a correctly performed two-sided null hypothesis test, sign errors are impossible in two-sided hypothesis tests. There are other tests where sign errors can't occur, such as an omnibus *F*-test used to detect differences between groups, as the *F*-distribution is based on squared deviations and only has positive values.

Researchers sometimes incorrectly make a directional claim after a two-sided test, which can lead to what Kaiser (1960) has referred to as a Type 3 error. Type 3 errors occur when H0 is false, a statistically significant effect is observed in the opposite direction of the true effect, and researchers make a directional claim. The correct way to make directional claims is to perform a one-sided test. If researchers want to make a claim about an effect in the positive or negative direction, they should follow Kaiser's (1960) proposal to perform two one-sided null-hypothesis tests, one in the positive direction, and one in the negative direction, each at α/2 (see also Leventhal & Huynh, 1996).

According to Gelman and Tuerlinckx (2000) effects are never exactly zero. This leads them to argue that it is uninteresting to perform a null-hypothesis test, or to control the Type 1 error rate. They propose to remove effects of 0 from the possible effect sizes that can be observed. This creates a situation where all effects are true effects, either in the

positive or the negative direction. A statistical test now has two possible results: Effects in the opposite direction as the true effect size are correctly rejected, or effects in the true direction are incorrectly rejected - a Type S error. Type S errors can be computed for any test that makes claims based on a threshold (for closed form formula, see Lu et al., 2019), but in this article we will focus on frequentist hypothesis tests. We set the alpha level for the test to 0.05 in all examples below.

If one believes effect sizes of 0 do not exist, it is no longer possible to perform a two-sided nill-null hypothesis test. Gelman and Tuerlinckx (2000) instead recommend to perform two simultaneous one-sided tests: One test in the positive direction, and one test in the negative direction. To understand Type S errors, it is useful to consider the most extreme scenario where the effect size is not zero. If there is an infinitesimally small positive effect, the statistical power of the test is practically indistinguishable from the alpha level of 0.05. Effects will be rejected in the positive direction with a probability of 0.025 in the long run (correct rejections), and effects will be rejected in the negative direction with a probability of 0.025 in the long run (Type S errors). The only reason we do not call a Type S error a Type 1 error is because Gelman and Tuerlinckx have removed the value of 0 from the distribution of possible effect sizes. A Type 1 error consists of all significant test results for values ≤ 0, while a Type S error consists of all significant test values < 0. As any point in a continuous distribution has 0 probability, excluding the value of 0 will not change the long-run probabilities. Therefore, a Type S error has the same probability as a Type 1 error in a one-sided test. The difference between a Type 1 error and Type S error is purely conceptual, as Type 1 and Type S error rates are identical in one-sided tests.

When two one-sided tests are performed and the null is never true, the Type S error rate is at most $\alpha/2$. Gelman and Turelinckx prefer to express the Type S error rate as a

proportion of the significant results and "believe that this conditional probability is the appropriate error rate to consider, since our primary concern is to understand the frequency properties of claims with confidence". In the infinitesimal effect case the statistical power of the test is 0.05 and the Type S error is 0.025, so the rate of Type S errors as a proportion of significant results is 0.025/0.05 = 0.5. In the extreme case where power is as low as the alpha level, 50% of significant results are sign errors.

This conditional probability is similar to the false discovery rate, which is the expected proportion of false positives among all positive findings. Just as Type 1 errors only occur when the null hypothesis is true, the false discovery rate only occurs when the null hypothesis is true. But once again, removing the value of 0, which itself has 0 probability, does not change the long-run false discovery rate. If researchers believe the null is true, a Type 1 error (a rejection of positive effects, when there is a positive effect) will be observed with a probability of 0.025, and a correct rejection (a rejection of negative effects, when there is a positive effect) will occur with a probability of 0.025. From all positive findings (0.025 false positives + 0.025 true positives = 0.05 positive test results) 50% are false positives. Therefore, the false discovery rate is 50% in the two one-sided testing procedure. The conditional Type S probability is identical to the false discovery rate, with the conceptual difference that it excludes the value of 0 from the hypothesized values.

Should effects of 0 be considered impossible?

Note that it is somewhat peculiar to completely exclude the possibility that effect sizes of 0 exist in a hypothesis test. After all, one could just as easily argue that in continuous data, no true effect size is exactly 0.5, but no one would propose to remove 0.5 from the space of hypothesized values. Researchers have also disagreed with the idea that effect sizes are never zero, with Krueger and Heck (2019) stating that: "for many questions humans ask of nature, the null (or any particular tested hypothesis) may in fact be true."

Frick (1995) concludes that "for some experiments, the null hypothesis is possible". And Hagen (1997) drives home the criticism even more forcefully: "If, as some have claimed, the null hypothesis is always false, we would be foolish, indeed, to spend time conducting statistical tests that can only tell us what we already know. But we need not feel foolish. As far as I can tell, the claim has never been sustained by either statistical or logical arguments". Neither side can provide empirical support for or against the hypothesis that the null hypothesis is never true, as we can't measure the entire population for all effects scientists want to study. The claim that the null is never true is scientifically unfalsifiable.

It seems difficult to justify why one would believe the null is never true, while there is no doubt that an effect size of 1×10^{-32} can be true. Instead of discussing which effect sizes can be true, a more sensible concern is the idea that all variables are connected through theoretically uninteresting causal structures that result in non-zero correlations between variables (especially in observational studies). In psychology, this idea is referred to as the crud factor (Meehl, 1990; Orben & Lakens, 2020). The crud factor has been one argument to move beyond null hypothesis significance tests, and instead test against a range of theoretically or practically uninteresting effect sizes around zero by performing a minimum-effect test, or a test for superiority (Lakens et al., 2018).

Instead of performing a test that rejects effect sizes of 0, researchers can specify a smallest effect size of interest, and test whether they can statistically reject all effects that are deemed too small to matter. Such a minimum-effect tests resolve most of the problems researchers have with null-hypothesis significance testing (Lakens, 2021), including the concern that the null is never true. If researchers do not believe it is scientifically interesting to reject effects of exactly 0 they can test whether effects in a range around 0 can be rejected. For example, Ferguson and Heene (2021) empirically show that correlational effects in aggression research are unlikely to ever be exactly 0, and established a correlation

of $r = 0.1$ are a lower bound for hypothesis tests in this research field. A directional minimum-effect test against $r \leq 0.1$ is rejected if the lower bound of the 90% confidence interval around the observed correlation is larger than 0.1 (Lakens et al., 2018). minimum-effect tests are a better solution than arbitrarily excluding a point value of 0 from effects that are possible. If minimum-effect tests are adopted, the idea of a Type S error is no longer needed, as it is equivalent to a Type 1 error for one-sided tests against a non-nill null hypothesis.

Should researchers report Type S errors in scientific articles?

Whereas Gelman and Tuerlinckx (2000) only discuss the relevance of Type S errors relative to Type 1 errors, Gelman and Carlin (2014) take an additional step and suggest that it is useful to compute and report Type S error rates for specific studies. To compute a Type S error researchers need to specify what they believe is the true effect size. Gelman and Carlin suggest using a literature review or other available data. They provide an example where the original author observed an effect of 8 percentage points. Gelman and Carlin do not believe this finding can be correct, and retrospectively compute the Type S error based on effect sizes they believe are more likely to be true: 0.1, 0.3, and 1 percentage point. When they compute the Type S error for the much smaller effects they deem plausible, the probability of a sign error can be quite large as long as the statistical power is low.

Is it useful to compute and report Type S errors in this way? One problem with retrospective design analysis is that if a skeptic wants to argue that an effect size is unreliable, a retrospective design analysis will practically never prove them wrong. For any effect where a skeptical reader feels the need to question the scientific claim by performing a retrospective design analysis, there will be substantial uncertainty about the true effect size. In these cases, the literature (or any other external data) will rarely - if ever - provide strong constraints on how small effect sizes can plausibly be. This means a skeptical reader

can always find reasons to perform a retrospective design analysis for very small effect sizes, which make it easy to claim that there is a high probability of a sign error. If the skeptic was wrong, a retrospective design analysis would practically never tell them they are wrong, and claims about high Type S error rates are therefore rarely severely tested.

To conclude this section, we have strong conceptual issues with Type S errors (removing 0 from the range of possible distributions is difficult to justify, and the long run probability of Type S errors is identical to the well-established Type 1 error rate and false discovery rate), as well as practical concerns (the test can too easily lead to a foregone conclusion). Type S errors mainly point out how uninformative studies with low statistical power are, and a null hypothesis test might not be an interesting question to ask. If researchers want to address these concerns, a better solution is to specify a non-null nill hypothesis (e.g., considering all effects between −0.1 and 0.1 as practically equivalent to 0) and perform an a-priori power analysis for the smallest effect size of interest. After the data is in, the Type S error is at most 2.5%, as is the Type 1 error. Researchers should always make a claim while acknowledging the maximum possible Type 1 error rate (e.g., 2.5% in an one-sided test with an alpha level of 0.025) and any quantification of how much lower the Type S error might be, depending on weakly informed guesses of the true effect size, comes with great uncertainty.

Gelman and Carlin (2014) argue that a retrospective design analysis can reveal that 'a study was too small to be informative'. But this seems misguided. If a significant effect was observed in a well-performed test we can reject the null with a maximum Type 1 error rate of 2.5%, which is small enough to take the result seriously. The study might have had very low power to detect an effect, but if it happened to detect an effect with the desired maximum Type 1 error rate, the fact that the study had a low probability to yield an informative outcome a-priori no longer matters. The observed effect size might be

measured inaccurately, and inflated due to selection bias (see Type M errors below) but a researcher can claim that there is a very small probability that random noise is mistaken as a true effect.

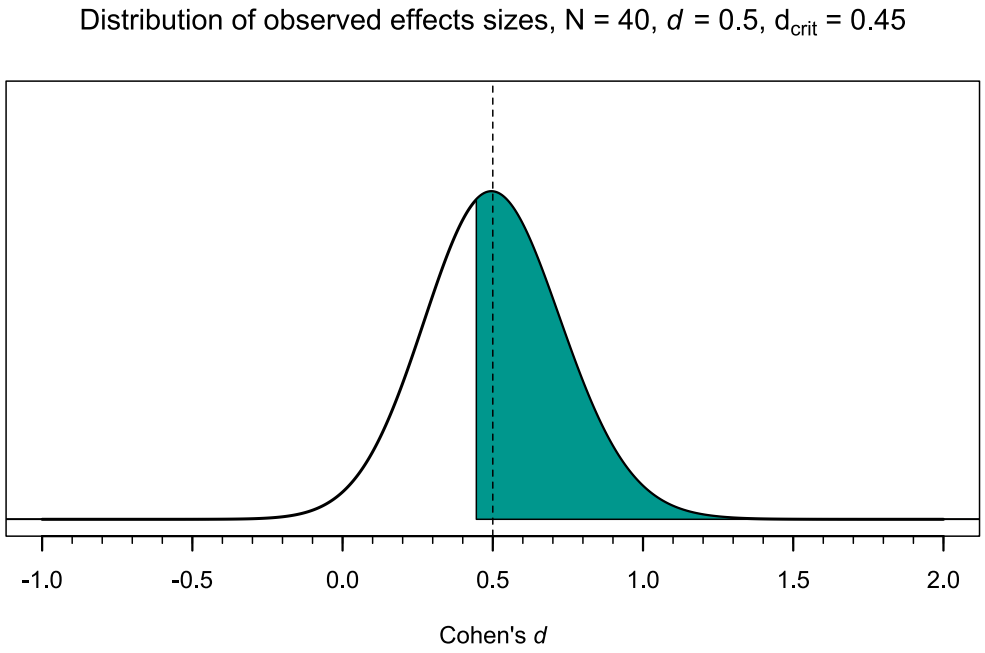
The Type S error rate will only raise a red flag for studies with incredibly low statistical power. When the Type S error is high, the design is uninformative, but when the Type S error is low, the design is often still uninformative. The Type S error drops to negligible levels when the statistical power of the test is still unacceptably low. For example, Simonsohn (2015) considers a test with less than 33% power severely underpowered, but with 33% power the probability that a significant result is in the wrong direction is only 0.1%.

Effect size inflation and Type M errors

Lane and Dunlap (1978) examined the impact of selection bias on the inflation of effect size estimates. When only statistically significant results are shared, the effect size estimate based on these studies is inflated, as smaller non-significant effects are removed from the scientific literature. In general, how much effect sizes are inflated in the presence of selection bias is a function of the power of the test. The lower the statistical power - for example because population effect sizes or sample sizes are smaller - the more significant effect sizes will be inflated. Figure 1 shows the expected effect size distribution for an alternative hypothesis with a true effect size of Cohen's *d* of 0.5. With a sample size of 40 participants in each group, a two-sided independent *t*-test will yield a statistically significant result if the observed effect size is larger than $d = 0.445$. If there is extreme selection bias for statistically significant results the scientific literature would consist of effect sizes to the right of $d = 0.445$. Instead of having access to the full distribution of expected effects, the literature will represent a truncated distribution, where all values smaller than 0.445 are removed from the distribution of expected effect sizes.

Figure 1

Expected effect size distribution under the alternative hypothesis, assuming a true effect size of 0.5 (dashed central line). With a sample size of 40 participants per group, a two-sided independent t-test yields a statistically significant result when the observed effect size exceeds d = 0.45. The shaded area represents the subset of effect sizes published under selection bias.



This type of bias has been discussed extensively in the literature (Anderson & Maxwell, 2017; Begg & Mazumdar, 1994; Hedges, 1984). Both (Hedges, 1984) as Taylor and Muller (Taylor & Muller, 1996) have developed a statistical model that computes bias-adjusted effect size estimates when estimators are selected from a 'censored' or 'truncated' distribution. This method creates a model of what the full distribution would have looked like, and estimates the effect size if there was no selection bias. Anderson, Kelley, and Maxwell (2017) apply this method to compute an adjusted estimate of the noncentrality parameter, and use it to show how a-priori power analysis can be adjusted for selection bias.

Gelman and Carlin (2014) propose another approach to increase awareness of effect size inflation due to selection bias. They compute a ratio between the average absolute effect size estimates from statistically significant effects, and the assumed true effect size, which they refer to as a Type M (magnitude) error. Because Type M errors do not quantify error rates (i.e., the probability of an incorrect claim) but the bias in an estimator, we prefer Gelman and Carlin's alternative term for Type M errors: the 'exaggeration ratio'. The exaggeration ratio is a function of the statistical power of a test. When power is close to 100% all tests of true effects will be significant, no effect sizes are removed due to selection bias, and the exaggeration ratio will be 1, indicating that the design will in the long run yield unbiased effect size estimates. As power is practically always less than 100%, the exaggeration ratio will typically be larger than 1, indicating that effect sizes selected for statistical significance will be inflated. Altoè et al. (2020) note that when power reaches at least 80% the average overestimation of the effect size tends to be just above 10%, which they consider practically negligible. As such, one may see the exaggeration ratio as a complementary perspective on the information conveyed by statistical power, shifting the focus from hypothesis testing to effect size estimation.

The exaggeration ratio is computed based on properties of the design of the study, and the assumed true effect size. The design (e.g., the sample size, type of test, and the alpha level) determine the critical effect size, or the smallest effect size that can be statistically significant. The assumed true effect size determines how inflated the average effect size in statistically significant results will be on average, with the smaller the assumed effect size, the larger the exaggeration ratio. The exaggeration ratio (Gelman & Carlin, 2014) is not intended to correct individual effect sizes. If the average absolute unbiased effect size estimate is 0.5, and the average biased effect size estimate is 1, the estimated effect is on average 2 times larger than the true effect size. But this exaggeration

ratio of 2 applies to the average effect size estimate, and not to any individual effect size. It is not correct to simply divide each observed effect size estimate by the exaggeration ratio and treat it as a bias-adjusted effect size estimate.

Despite the fact that the authors never intended the exaggeration ratio to be used to adjust individual effect sizes, researchers have misused Type M errors for this purpose. In the following example, Shem-Tov et al. (2024) note that the Type M error rate quantifies an average inflation, but still misuse it to adjust the observed effect size: "following the procedure proposed by Gelman and Carlin (2014), we estimate an average potential exaggeration ratio of 1.2 in the effect of enrollment to MIR on rearrests within one and four years. In other words, on average, our estimates might indicate that the impact of enrollment to MIR causes a reduction of 23.4 percentage points while the true effect is a reduction of 19.5 percentage points." Similarly, Gajendran et al., (2022) write: "the modest Type M exaggeration ratio of 1.27 indicates the possibility that the communication medium effect is overestimated by a factor of 1.27, which is inconsistent with the effect being an unlikely result." However, the average inflation is not the same as the inflation in any individual study, and the inflation might be much larger or smaller for this specific study.

There are statistical approaches that adjust effect size estimates for selection bias, such as likelihood based approaches (Hedges, 1984; Taylor & Muller, 1996) and *p*-uniform (Assen et al., 2015). The approaches developed by Hedges (1984) and Taylor and Muller (Taylor & Muller, 1996) to adjust effect size estimates for selection bias takes the observed effect size as a biased estimator, and based on a model for selection bias, computes the maximum likelihood estimate for the effect size after correcting for bias. Anderson and colleagues (2017) have implemented this bias correction method in the BUCSS R package, and extended the approach for normally distributed data to other statistical distributions. Although most other bias correction methods can only be used meta-analytically (i.e., they

need multiple effect sizes to correct for bias) the *p*-uniform technique developed by Van Assen and colleagues (2015) uses a similar model for selection bias as Hedges (1984), and can be performed on a single study. Note that BUCSS is aimed at sample size calculations, and does not compute bias-corrected effect sizes directly, but the non-centrality parameter that the function returns can be used to compute a bias-corrected effect size estimate. These approaches (as the simulations below will show) allow researchers to adjust observed effect sizes for bias corrected effect sizes.

In addition to bias-corrected effect size estimates, researchers can also report the critical effect size (Perugini et al., 2025). The critical effect size is the smallest effect size that could reach statistical significance. Just as the Type M error, it is computed based on the study design, and not the study results. The critical effect size similarly increases awareness about how selection bias inflates effect size estimated from studies selected for statistical significance, but it also warns against interpreting non-significant results as the absence of an effect by pointing out the range of non-zero effect sizes that would never yield a statistically significant result. Critical effect sizes are also worth reporting in studies with very high statistical power, where the Type M error would not be reported because it is negligible. In studies with extremely high power, the critical effect size will make it clear that even trivially small effects will be statistically significant. This can make researchers reflect on which effects are large enough to be meaningful.

Simulating effect size inflation and correction methods

We simulated 10.000 independent *t*-tests with a true effect size of $d = 0.5$, 40 observations per group, equal variances, and an alpha level of 0.05. We also computed adjusted effect size estimates using *p*-uniform and the bias and uncertainty-corrected sample size procedure (BUCSS). Note that the BUCSS R package returns a non-centrality parameter, which was transformed into a bias-adjusted effect size estimate. Van Aert et al.

(2019, p. 16) note how the likelihood-based procedure implemented by Anderson and colleagues (2017) is 'based on similar methodology' as *p*-uniform. The bias and uncertainty-corrected sample size procedure (BUCSS) by Anderson and colleagues does not return an adjusted value if the effect size estimate is negative. *P*-uniform does return negative estimates. When applying *p*-uniform in its intended context of meta-analyses, the recommendation is to replace negative estimates by zero (Aert et al., 2019). This solution is less ideal when analyzing single studies because too many estimates will be set to 0, which negatively biases the estimate. Instead, we follow Anderson and colleagues (2017) and remove negative estimates. Although we do not think this has been pointed out before in the literature, the corrections based on BUCSS and *p*-uniform are practically identical.

As noted above, the exaggeration ratio should not be used to adjust individual effect sizes, and was not developed for such a use. The distribution in the top-right pane of Figure 2 shows why this approach fails. Large observed effects are adjusted downward a lot, but smaller observed effect sizes are adjusted downward less, and not sufficiently. Where the average inflation can be used to educate researchers about the risk of inflation, it can't be used to adjust individual effect sizes for selection bias.

Figure 2 shows the effect size distribution of observed effect sizes when only statistically significant results are available (top-left pane). The total unbiased dataset would lead to a symmetrical distribution of observed effect sizes around 0.5. Given the sample size, only effects larger than $d = 0.445$ will be statistically significant, and all effects smaller than $d = 0.445$ are missing from the truncated distribution. The BUCSS and *p*-uniform methods lead to practically identical distributions of adjusted effect size estimates (bottom panes). The adjusted effect size estimates do not perfectly match the unbiased distribution, but provide estimates that are on average much closer to the true effect size. This demonstrated that it is more useful to report bias-adjusted effect size estimates using

for example the *p*-uniform technique, than to report the Type M error - or even worse, misuse the exaggeration ratio to adjust observed effect sizes.

Figure 2

Comparison of the observed and bias-corrected effect size yielded by BUCSS, p-uniform, and the exaggeration ratio, and the statistical power of the test.

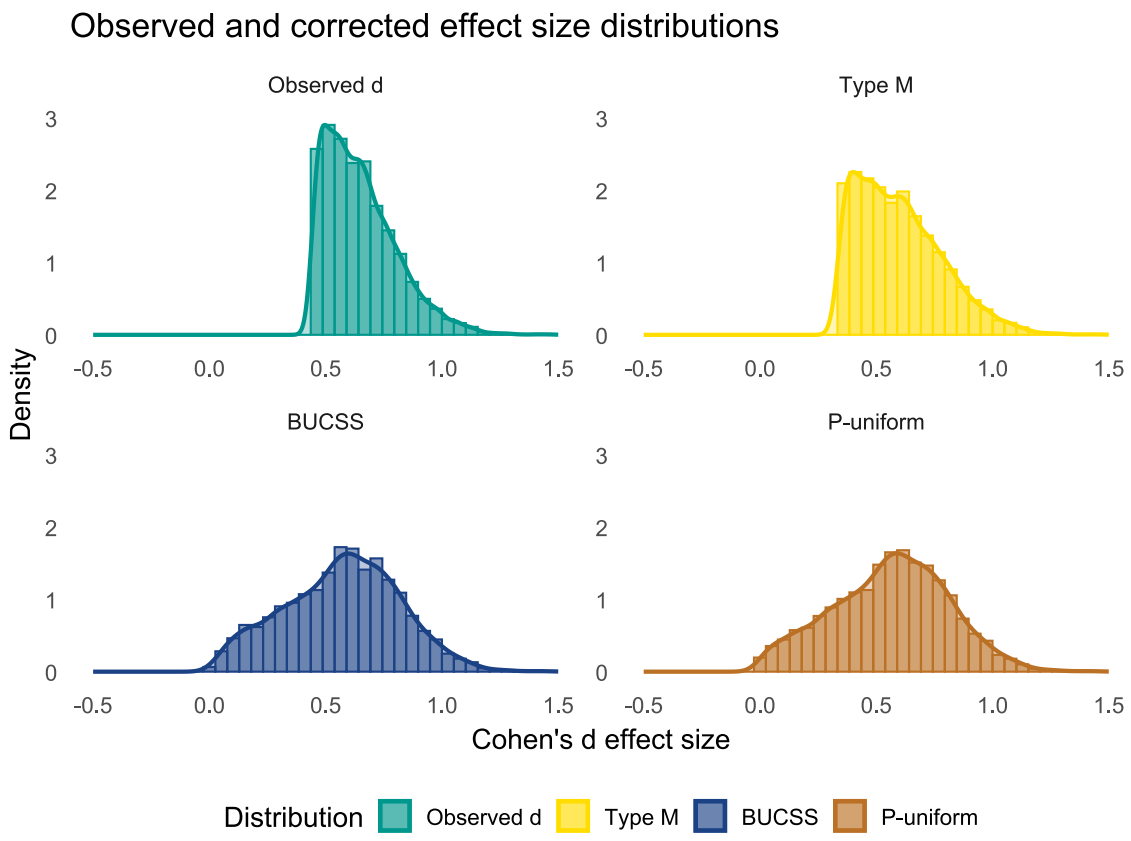
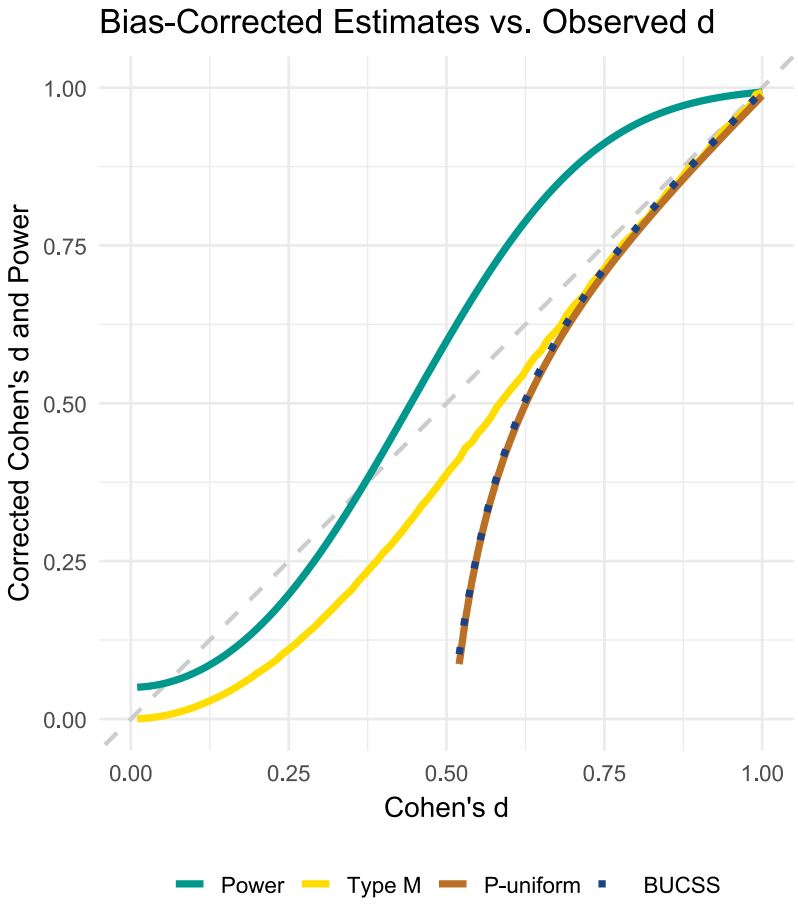


Figure 3 plots the observed Cohen's *d* and bias-corrected Cohen's *d* for independent *t*-tests with true effects from *d* = 0 to *d* = 1, 40 participants per group, and an alpha level of 0.05, with observed effect sizes on the y-axis, and corrected effect size estimates on the x-axis. The power of the test (which is a function of the effect size) is also plotted, and we see that when power is high (when Cohen's *d* is large, e.g., from *d* = 0.75 onward) the observed Cohen's *d* is practically equal to the corrected Cohen's *d*. This is because there is very little

selection bias, as most tests yield statistically significant results, and the observed effect size is not inflated.

Figure 3

Comparison of the corrected effect sizes for observed Cohen's d effect sizes from 0 to 1 using BUCSS, p-uniform, and Type M errors.



As the observed effect size becomes smaller, power is lower, selection bias is larger, and the corrected effect sizes become much smaller than the observed effect sizes for the BUCSS and *p*-uniform methods. We again see that these two methods yield basically the same corrected effect sizes. In this testing scenario the critical effect size is $d = 0.445$, which means effects smaller than this value will not yield a statistically significant result, and the correction methods only start to estimate positive effects from observed effect sizes of $d =$

0.52, as can be seen in the plot. Again, the yellow line visualizing a correction based on the exaggeration ratio illustrates why it should not be used to correct individual effects.

The exaggeration ratio assumes extreme selection bias where only statistically significant effects are available, which might be a reasonable expectation in some literatures (e.g., Scheel et al., 2021), while there are more non-significant results published in other other literatures (e.g., Mesquida et al., 2023). The only way to ensure unbiased effect size estimates is for researchers to report all results that are obtained, regardless of the significance level of test results. This can be achieved by reporting all studies in a public registry, or by publishing research as a Registered Report.

The Use of Type S and M Error in an "Imperfect" Science

Although Type S and M errors might not be particularly informative to report in a result section, given the more informative alternatives (i.e., minimum-effect tests and bias-corrected effect size estimates), they may still be valuable as a tool to reflect on the problems that emerge when studies are underpowered and/or selectively reported. Due to resource constraints, researchers may sometimes perform studies with very low power for the main effect of interest. Consider a PhD student who can at most collect 20 participants per group for an independent two-sided *t*-test, and expects a true effect size of $d = 0.4$. An a priori power analysis reveals that the statistical power is a mere 23%. With such a high probability of uninformative results, one might argue that the study should not be performed. However, the study is part of a grant proposal that funds the research of the PhD student, and their supervisor wants to complete the promised data collection, regardless of the low statistical power. Under these suboptimal conditions the student could reflect on the Type S and M error rates in a preregistration. They could point out that the effect size estimate has high uncertainty, that there is a probability of 0.3%. that a statistically significant effect is in the wrong direction. Perhaps more informatively, the PhD

includes the Type M error, warning readers that statistically significant effects will on average be inflated with a ratio of 2.06, meaning that on average any significant effect would be overestimated by 100%. This should make researchers aware of the fact that due to the underpowered nature of the study, the effect size estimate can not be taken at face value, especially if the effect size will only be interpreted if it is statistically significant.

A second use-case of a Type S and M error is in highly exploratory research where researchers opportunistically search for statistically significant effects without correcting for multiple comparisons. In such a scenario true effects might be small for some tests, researchers often selectively focus on statistically significant effects, and in the absence of strong theoretical predictions a skeptical peer might argue true effects might be small. Type S and M errors might be a way to communicate that exploratory analyses can be uninformative, and an exploratory search for significant effects will on average lead to exaggerated effect size estimates. This might help researchers realize that they should not make claims based on exploratory analyses as error rates can be high. Instead, results from exploratory analyses should be treated as hypotheses that need to be severely tested in follow-up studies (Ditroilo et al., 2025).

The Use of Type S and M Errors in Statistics Education

Statistical misconceptions are widespread among researchers. We believe that some statistical misconceptions that we have heard can be mitigated by educating researchers about Type S and M errors. For example, the statement ‘If the power of a study is low, the main risk is failing to detect an effect that is actually present’ overlooks the risk that a statistically significant effect from an underpowered study may be substantially overestimated. Another common misconception is that “If the sample is small and the result is statistically significant, the effect must be large”, which fails to acknowledge that the effect can be small, and that all effects selected for significance are substantially inflated.

Type S and M errors can serve as effective educational tools to challenge such misconceptions and promote a deeper reflection on the risks involved in statistical inference under conditions of low power and selective reporting.

The idea of a Type S error can also help students grasp why it is incorrect to make directional claims after a two-sided test (Cho & Abe, 2013). The notion of a Type S error provides an opportunity to discuss why one-sided tests are necessary for directional hypotheses, and to introduce the practice of two one-sided tests at $\alpha/2$ to test effects in both directions (Kaiser, 1960; Leventhal & Huynh, 1996). Rather than teaching Type S errors as statistical quantities to compute and report in a manuscript, instructors can use them to emphasize the importance of aligning statistical decisions with the inferential goals. Instructors may also introduce the idea that when power is very low, even the direction of a statistically significant result can be misleading. This message may resonate more with students than visualizing power curves.

The concept of a Type M error—or exaggeration ratio—is especially helpful when introducing students to the consequences of publication bias and selective reporting. Even without delving into the mathematical derivation, Figure 1 clearly shows how filtering for significance results in inflated effect size estimates. This can serve as a foundation for teaching the idea that significant effects from underpowered studies are not only uncertain, but often systematically overestimated. The same idea can be taught through the concept of a critical effect size (Perugini et al., 2025) by illustrating how low power limits which effect sizes can be reliably distinguished from random noise. In follow-up courses, teachers could introduce the exaggeration ratio as a function of the statistical power of the test, and explain the limitations of underpowered studies. The Type M error can also be used to explain the difference between the inferential goals of hypothesis testing and estimation, revealing that tests that reject the null hypothesis inform us about the presence of an effect,

but do not provide accurate effect size estimates. In more advanced courses, instructors may introduce tools for bias detection (e.g., funnel plots, p-curve analysis) or methods to compute bias-adjusted effect size estimates (Assen et al., 2015; Bartoš & Schimmack, 2020; Simonsohn et al., 2014; Stanley, 2017).

When taught properly, the concepts of Type S and M errors can help to bridge the gap between statistical theory and the realities of scientific practice. They offer a narrative that aligns with the goals of open and rigorous science: understanding the risks of false directional claims, exploratory tests without error control, underpowered studies, and bias in the published literature, and emphasizing the value of reporting all research findings, for example through study registries or Registered Reports. Teaching students to think about the inferential claims they can—and cannot— make is important to improve their statistical literacy, and will bolster their critical thinking skills when they read claims in the scientific literature.

Conclusion

Gelman and Carlin (2014) propose to compute Type S and M errors for planned or performed statistical tests (see also Altoè et al. (2020)). However, both for conceptual and practical reasons we see limited value in quantifying Type S errors and the exaggeration ratio (or Type M error) for individual studies. We share the general concern that researchers should design informative studies, and be cautious when interpreting the results from underpowered studies, especially when combined with selection bias. Type S and Type M errors can be used to create awareness of the limitations of studies where researchers did not follow best practices, and they can play a role in statistics education to improve students' understanding of how uninformative underpowered studies are. However, we believe there are more useful alternative statistical approaches to address these concerns. Instead of reporting Type S error rates, researchers should perform tests

against a range of values considered theoretically or practically equivalent to 0. Instead of reporting Type M errors, researchers should report bias-corrected effect size estimates provided by methods such as *p*-uniform, and report the critical effect size.

References

Aert, R. C. M. van, & Assen, M. A. L. M. van. (2018). *Correcting for publication bias in a meta-analysis with the p-uniform* method*. https://doi.org/10.31222/osf.io/zqjr9

Aert, R. C. M. van, Wicherts, J. M., & Assen, M. A. L. M. van. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

Albers, C. J., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02893

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 1–20. https://doi.org/10.1080/00273171.2017.1289361

Assen, M. A. L. M. van, Aert, R. C. M. van, & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

Bartoš, F., & Schimmack, U. (2020). *Z-curve.2.0: Estimating replication rates and discovery rates*. https://doi.org/10.31234/osf.io/urgtn

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101.

Bertoldo, G., Callegher, C. Z., & Altoè, G. (2022). Designing studies and evaluating research results: Type m and type s errors for pearson correlation coefficient. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2573

Callegher, C., Bertoldo, G., Toffalini, E., Vesely, A., Andreella, A., Pastore, M., & Altoè, G. (2021). *PRDA: An r package for prospective and retrospective design analysis*. https://doi.org/10.21105/joss.02810

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Ditroilo, M., Mesquida, Abt, & Lakens, D. and. (2025). Exploratory research in sport and exercise science: Perceptions, challenges, and recommendations. *Journal of Sports Sciences*, *43*(12), 1108–1120. https://doi.org/10.1080/02640414.2025.2486871

Ferguson, C. J., & Heene, M. (2021). Providing a lower-bound estimate for psychology's "crud factor": The case of aggression. *Professional Psychology: Research and Practice*, *52*(6), 620–626. https://doi.org/https://doi.org/10.1037/pro0000386

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*(1), 132–138. https://doi.org/10.3758/BF03210562

Gajendran, R. S., Loewenstein, J., Choi, H., & Ozgen, S. (2022). Hidden costs of text-based electronic communication on complex reasoning tasks: Motivation maintenance and impaired downstream performance. *Organizational Behavior and Human Decision Processes*, *169*, 104130. https://doi.org/10.1016/j.obhdp.2022.104130

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641651. https://doi.org/https://doi.org/10.1177/1745691614551642

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390. https://doi.org/10.1007/s001800000040

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *The American Psychologist*, *52*(1), 15–24. http://cat.inist.fr/?aModele=afficheN&cpsidt=10561120

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85. https://doi.org/10.3102/10769986009001061

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, *16*(2), 261–268. https://doi.org/https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*(3), 160–167. https://doi.org/https://doi.org/10.1037/h0047595

Krueger, J. I., & Heck, P. R. (2019). Putting the p-value in its place. *The American Statistician*, *73*(sup1), 122–128. https://doi.org/10.1080/00031305.2018.1470033

Lakens, D. (2020). Pandemic researchers — recruit your own best critics. *Nature*, *581*(78077807), 121–121. https://doi.org/10.1038/d41586-020-01392-8

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Leventhal, L., & Huynh, C.-L. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods*, *1*(3), 278–292. https://doi.org/http://dx.doi.org.dianus.libr.tue.nl/10.1037/1082-989X.1.3.278

Lu, J., Qiu, Y., & Deng, A. (2019). A note on type s/m errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 1–17. https://doi.org/10.1111/bmsp.12132

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/s15327965pli0102_1

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2023). Publication bias, statistical power and reporting practices in the journal of sports sciences: Potential barriers to replicability. *Journal of Sports Sciences*, 1–11. https://doi.org/10.1080/02640414.2023.2269357

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641650. https://doi.org/https://doi.org/10.1177/001316446002000401

Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Perugini, A., Toffalini, E., Gambarota, F., Lakens, D., Pastore, M., Finos, L., & Altoè, G. (2025). The benefits of reporting critical effect size values. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/7qe92

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. https://doi.org/10.1037/0003-066X.40.1.73

Shem-Tov, Y., Raphael, S., & Skog, A. (2024). Can restorative justice conferencing reduce recidivism? Evidence from the make-it-right program. *Econometrica*, *92*(1), 61–78. https://doi.org/10.3982/ECTA20996

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/https://doi.org/10.1037/a0033242

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 194855061769306. https://doi.org/10.1177/1948550617693062

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics-Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787