UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Sotola, L. K. (2022). Garbage In, Garbage Out? Evaluating the Evidentiary Value of Published Meta-analyses Using Z-Curve Analysis. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.32571

Methodology and Research Practice

Garbage In, Garbage Out? Evaluating the Evidentiary Value of Published Meta-analyses Using Z-Curve Analysis

Lukas K. Sotola[1] [a]

[1] Psychology, Iowa State University, Ames, IA , USA

Keywords: z-curve, replicability, meta-analysis, questionable research practices, publication bias, statistical power

https://doi.org/10.1525/collabra.32571

Collabra: Psychology

Vol. 8, Issue 1, 2022

The purpose of the current work was to examine the evidentiary value of the studies that have been included in published meta-analyses as a way of investigating the evidentiary value of the meta-analyses themselves. The studies included in 25 meta-analyses published in the last 10 years in *Psychological Bulletin* that investigated experimental mean differences were z-curved. Z-curve is a meta-analytic technique that allows one to estimate the predicted replicability, average power, publication bias, and false discovery rate of a population of studies. The results of the z-curves estimated a substantial file drawer in three-quarters of the meta-analyses; and in one-third of the meta-analyses, up to half of the studies are not expected to replicate and up to one-fifth of the studies included could be false positives. Possible reasons for these findings are discussed, and caution in interpreting published meta-analyses is recommended.

Meta-analysis has become an important tool in psychological science. Meta-analyses allow researchers to estimate the true effect size of an intervention or effect (Glass, 1976), to explore moderators of effect sizes, and to estimate the amount of variability in effect sizes that cannot be explained by methodological artifacts or examined moderators. Meta-analyses also allow researchers to suggest areas that require further research in their respective disciplines, and the average effect size estimates they generate can be used for power analyses in primary research (Sharpe & Poets, 2020). As such, published meta-analyses are often popular and influential, because they offer both a summary of what is known about a relationship or an effect, and insights into the types of future work in a domain that may be most valuable.

Despite its popularity and influence, meta-analysis as a method has received criticism from the moment it came into widespread use (Eysenck, 1978). While criticisms of meta-analysis are diverse (e.g., Cafri et al., 2010; Chevret et al., 2018; Ioannidis, 2016; Lakens et al., 2017), the most important criticism of meta-analysis is that the result of a meta-analysis will only be as reliable as the studies that go into it (Nelson et al., 2018). Biases in the mechanisms by which the data included in a meta-analysis are generated—whether it is due to methodological flaws in the primary studies, publication bias, researcher degrees of freedom, or a combination of these factors—will produce biases in the outcome of the meta-analysis. It is the purpose of this study to estimate the degree to which biases in data-generating mechanisms are present in recent influential meta-analytic syntheses of experimental research in psychology.

Up to this point, meta-analysts have dealt extensively with one kind of bias in data-generating mechanisms: publication bias (Ferguson & Brannick, 2012). Publication bias can lead to a population of nonsignificant results that do not get included in meta-analyses, which can inflate meta-analytic effect size estimates (Friese & Frankenbach, 2020). A number of methods for testing and correcting for publication bias have been developed and widely used (e.g., Duval & Tweedie, 2000; Orwin, 1983; Pustejovsky & Rodgers, 2019; Sterne et al., 2005). These methods all have their virtues and drawbacks (e.g., Carter et al., 2019), but they all attempt to provide an estimate of the degree to which publication bias may have distorted a meta-analytic effect size estimate.

However, there are biases in data-generating mechanisms that can distort meta-analytic effect size estimates that these tests for publication bias do not assess. In particular, they do not evaluate two phenomena: the presence of questionable research practices (QRPs) and the replicability of published studies. QRPs are a variety of practices that involve undisclosed flexibility with which researchers collect, analyze, and report their data. Examples include halting data collection once one finds statistical significance, selecting from among a few dependent variables for the one that yields a significant finding, and excluding (or including) outliers because doing so causes an analysis to yield a significant result (Simmons et al., 2011). The examples listed here are all instances of a specific kind of QRP known

a lksotola@iastate.edu

as p-hacking. QRPs, and p-hacking in particular, can increase one's chances of finding a false positive in an individual study, but, crucially, they can also bias the outcome of meta-analyses. One simulation study showed that if one meta-analyzed 100 experiments that had been done with small sample sizes (e.g., N = 20) and with a degree of QRPs, the result could suggest an average Cohen's *d* of around .5—even when the true effect was zero (Bakker et al., 2012; also see Friese & Frankenbach, 2020; Olsson-Collentine et al., 2021). It is even possible that QRPs bias the outcome of some of the tests for publication bias that many meta-analysts rely on (Carter et al., 2019). This makes QRPs an important bias in data-generating mechanisms for meta-analysts to evaluate, especially given that researchers admit to engaging frequently in these exact QRPs (John et al., 2012).

The replicability of the studies included in a meta-analysis is another important bias in data-generating mechanisms (Stanley et al., 2018). One review of 200 meta-analyses published in *Psychological Bulletin* found that the primary studies included in those meta-analyses had, at best, only 36% statistical power on average, and, at best, just under 10% of the studies included showed statistical power of 80% or higher. That suggests, in theory at least, that only 36% of the studies included in those meta-analyses would replicate if an exact replication were attempted (Stanley et al., 2018). Another study reviewed 15 published meta-analyses for which published primary multi-lab replication studies existed on the same effects of interest. The authors found that in only seven cases did the outcome of the replication study and the outcome of the meta-analysis align, meaning that the replication study and meta-analysis both suggested statistically significant effects in the predicted direction (Kvarven et al., 2020). In seven other cases, the meta-analysis found a significant effect, while the replication study did not (Kvarven et al., 2020).

How one should interpret a meta-analysis based on primary studies that do not replicate or are predicted not to replicate is up for discussion. However, one can probably tentatively assert that the results of such a meta-analysis should be interpreted with caution. Hence, meta-analysts should consider the possibility that their work will be distorted by the inclusion of primary studies that are not replicable. That said, it is not practical to expect replication studies to exist for meta-analysts to evaluate this in every case, nor is it feasible to expect meta-analysts to perform several replication attempts themselves before undertaking a meta-analysis. A method is needed to estimate replicability without actually engaging in replication studies.

And that is where z-curve comes in. In recent years, both p-curve analysis and z-curve analysis (Brunner & Schimmack, 2020; Simonsohn et al., 2014), have been developed to help estimate the degree to which researchers in an area may have engaged in some kinds of QRPs—in particular, p-hacking—and the predicted replicability of the studies in an area. It should be noted that both p-curve and z-curve can detect the possible presence of p-hacking, but cannot pick up instances of QRPs where statistically significant findings are not reported reliably in the published literature. Both methods rely on the insight that results obtained through p-hacking should be characterized by an unusually large proportion of p-values that fall between .01 and .05. In order to demonstrate why p-values between .01 and .05 should be relatively rare, consider the following three situations in which researchers study a difference between two groups with a real effect size of $d = .3$ using a two-sided test, and an alpha level of .05.

In the first situation, assume a population of studies that are all characterized by a sample size of 289, which suggests statistical power of 95%. Under this scenario, 95% of all observed p-values will be less than .05. However, a sample size of 289 also gives approximately 85% statistical power when the alpha level is .01. In other words, 85% of the observed p-values will be less than .01. Put yet another way, almost 90% of all p-values that are statistically significant at alpha = .05 will have p-values less than .01, making p-values of less than .01 approximately 9 times as likely as p-values between .01 and .05 (Sellke et al., 2001). In this first situation, even if authors only submit or journals only publish statistically significant results, then a histogram of the p-values in the literature will be characterized by a heavily positively skewed distribution of p-values. That is, most p-values will fall under .01, with a few above .01. A meta-analysis of effect sizes in this first situation would only show a small amount of bias, with 5% of non-significant (given 95% power at the .05 level) results not obtained by the meta-analysts.

Now consider a second situation where researchers study the same phenomenon, this time with a sample size of 86 such that the level of statistical power is only 50%. That is, only 50% of observed p-values will be less than .05. Under these same conditions, a sample size of 86 also provides a level of statistical power of approximately 24% if alpha were set to .01 rather than .05. In other words, approximately half of the observed p-values that fall below .05 would be less than .01 and approximately half would fall between .01 and .05. If authors only submit or journals only publish statistically significant results, then the literature will be characterized by a relatively uniform distribution of p-values. That is, a histogram of p-values below .05 would be nearly flat. Thus, a meta-analysis of effect sizes in this second situation would show heavy bias insofar as meta-analysts would not obtain approximately 50% of non-significant results.

Finally, consider a third situation where researchers in a single kind of p-hacking: they conduct their statistical comparison of the two groups at the end of every day's data collection and decide to stop data collection once they obtain a p-value of less than .05. In this situation, the histogram of p-values is likely to exhibit a negative skew, with most p-values falling just below .05 and only very few falling below .01 (Simonsohn et al., 2014). This process produces inflated effect size estimates (see Bakker et al., 2012), so a meta-analysis of this literature would be likely also to produce an inflated effect size estimate.

Both p-curve and z-curve analyze the p-values taken from a set of studies of interest to predict biases in data-generating mechanisms and the replicability of the studies included. They both compare the observed and expected distribution of p-values to determine the prevalence of p-hacking and replicability for a set of published study. The major difference between the two methods is that p-curve uses p-values directly, while z-curve converts p-values into

two-tailed z-statistics and analyzes the z-statistics. Both provide estimates of replicability and average power of the studies entered into the analysis, and estimate the degree of publication bias. While p-curve has become relatively popular in recent years, z-curve analysis appears to perform better under conditions of effect size heterogeneity (Brunner & Schimmack, 2020). Effect size heterogeneity is, of course, fairly common in meta-analyses (Stanley et al., 2018), so z-curve is relied on in this study. Thus, in the current work, z-curve will be used to estimate the evidentiary value of studies that have been included in published meta-analyses as a way of estimating the evidentiary value of the meta-analyses themselves.

Method

Pre-Registered Method

The methodology and analytic plan for this study were pre-registered and can be found, along with the R code and a reference list of all of the meta-analyses included in this review, at the Open Science Framework (OSF) page at this link: https://osf.io/gr3ax/?view_only=49a2ad228222477abe996cc627ce69b5. There are a few changes to note from this pre-registered method. First, in the pre-registered Method, it says that all of the meta-analyses coded for this project would be uploaded as PDFs on the OSF. However, it was decided not to do this, because that might end up infringing on copyright laws for those publications. Second, in the pre-registered Method, it is indicated that 34 meta-analyses were gathered for inclusion; however, only 21 were included in the final analysis. This is either because the z-curve analysis did not run (see Results for an explanation) or because the meta-analyses were found not to be suitable for inclusion upon closer inspection. See Inclusion Criteria below for more details on that.

The last difference between the pre-registered Method and the current Method is in how p-values were computed. The equations included below for how p-values were computed from effect sizes differ slightly from those in the pre-registered Method, because some further research on how to compute p-values from effect sizes was done, and a more valid way of doing so was found while double-checking the coding. Specifically, some coding had to be re-done in cases where studies included in meta-analyses manipulated their independent variables within-subjects, as the process for computing a p-value from a mean difference effect size (e.g., Cohen's *d*) is slightly different depending on whether the independent variable in the primary study was manipulated within- or between-subjects (see Coding Procedure for more details on this).

Literature Search

Psychological Bulletin was chosen as the target journal, because it represents one of the most prestigious outlets for meta-analyses and because it publishes many meta-analyses. The time interval of 2010 to the present (i.e., spring 2020) was chosen so that the meta-analyses included would be reasonably current. Only meta-analyses that investigated experimental mean differences between two groups, a treatment and control, were included. This was done because the p-values taken from the effect sizes included in such meta-analyses were most likely to fulfill an important assumption of z-curve: that the p-values being entered into the z-curve analysis are the p-values from the statistical tests for the *main hypothesis* of each study. This is crucial to z-curve analysis, because it is the p-values from the test of the central hypothesis that are likely to be subject to publication bias (Simonsohn et al., 2014). This assumption is much more likely to be met in the case of p-values taken from the effect sizes included in meta-analyses of comparisons between two groups rather than meta-analyses of correlational effects, because correlational meta-analyses often include correlation coefficients that were only peripherally related to the main hypotheses of the studies from which they were taken. A comparison between two groups is much more likely to be the *central test* of a study.

Inclusion Criteria

Only meta-analytic reviews of experimental literatures that included tables with the authors' coding for effect sizes and sample sizes, either in the main body of the article or in supplemental materials, were included in this review. The table also had to indicate whether each study came from a between- or within-subjects design, or the paper itself had to make it clear at some point in the text that 100% of studies were either between- or within-subjects designs.

According to the database PsycINFO, 568 pieces were published in *Psychological Bulletin* between January 1st, 2010, and December 31st, 2020. A total of 158 were meta-analyses (others were corrections, commentaries, and other kinds of pieces). Thirty-four were found to be tests of mean differences and to have tables from which effect sizes could be extracted, and were examined more in-depth for inclusion. From among the 34, a number were excluded because it was not clear which studies included were between-subjects or within-subjects. This information was necessary in order for the effect sizes to be properly converted to p-values. Others were excluded because they turned out not to be meta-analyses of mean differences manipulated experimentally, but were tests of quasi-experiments or cross-sectional studies (e.g., gender differences). Twenty-three meta-analyses ended up being coded and analyzed, although only 21 yielded data where a z-curve could be run successfully (see Results).

Coding Procedure

Extracting effect sizes

The major piece of coding was extracting the p-values from each individual effect size included in each meta-analysis. Because only meta-analyses that report their coding in a table of some sort were included in this project, this involved examining the effect sizes and sample sizes from these tables and computing test statistics—usually t-scores—from those values, and then generating p-values based on those test statistics. In a single case (Hu et al., 2020), the authors provided p-values alongside the effect sizes in their table, so these p-values were used for the z-curve in that case. In a second case (Weingarten et al., 2016), the authors performed a p-curve themselves and in-

cluded the disclosure table in their supplemental materials with all of the test statistics from the studies included, so in this case, the p-values from their disclosure table were used for the z-curve.

Calculating p-values

P-values were calculated differently depending on how the authors of each meta-analysis presented their results, what effect size they reported, and the designs of the studies they included. In three cases (Fox et al., 2011; Landau et al., 2015; Sedlmeier et al., 2012), the authors reported effect sizes as Pearson *r* coefficients, even though the research area was in large part experimental. In these cases, the Pearson coefficients were entered into the p-value generator along with the sample size to acquire p-values at this link: https://www.socscistatistics.com/pvalues/pearsondistribution.aspx.

For all other meta-analyses included, calculating the p-values was more complicated. This procedure was different depending on whether the design of each study included in each meta-analysis used a between- or within-subjects design (Hedges, 1983). The design of included studies was typically noted explicitly in the text. For example, Fischer et al. (2011) noted that all included studies had a between-subjects design, while Cracco et al. (2018) noted that all included studies had a within-subjects design. Some authors (e.g., Coles et al., 2019) noted that the included studies were a mix of between- and within-subjects designs, and for these meta-analyses, the precise design for each included study was typically included in the meta-analytic table that listed the effect sizes taken from each sample. In these cases, the p-values from the studies listed as using between-subjects designs and the studies listed as using within-subjects designs were calculated differently according to the processes described below.

Converting the effect sizes into p-values was different depending on whether the authors presented their effect sizes as Cohen's *d*, Cohen's *dz* (d from a within-subjects design), or Hedges' *g*. To compute p-values for between-subjects designs, it was often necessary first to convert Hedges' *g* back into Cohen's *d*. If the authors of the meta-analysis did not specify what method they used to convert Cohen's *d* into Hedges' *g*, the basic formula for the conversion given by Hedges (1983) was used: $d = g/c(m)$, where $c(m) = 1 - (3/((4 * m) - 1))$, where m is the degrees of freedom for the study. m was calculated as $N - 2$ for studies that used between-subjects designs and $N - 1$ for studies that used within-subjects designs. If the authors specified that they performed the conversion from Cohen's *d* to Hedges' *g* another way, then this method was used. If Cohen's *d* was what the authors reported in their meta-analytic table, then no conversion was necessary and the Cohen's *d* could be converted directly into a t-score.

Once a Cohen's *d* was obtained, it was converted into a *t*-score. If the t-score was from a between-subjects design *and* sample sizes for the treatment and control conditions were provided in the meta-analytic table, it was converted to a t-score using the following formula: $t = d/\sqrt{(\frac{1}{n1}) + (\frac{1}{n2})}$. If the sample sizes for the treatment and control groups were *not* provided in the meta-analytic table, the t-score was calculated as follows: $t \approx \frac{d}{2} * \sqrt{N}$. The degrees of freedom for each t-score was calculated as $N - 2$. In the case of within-subjects designs, the Cohen's *dz* was converted to a *t*-score using the following formula: $t = dz * \sqrt{n}$. The degrees of freedom was then calculated as $N - 1$. All t-scores were converted to p-values at the following link, assuming a two-tailed hypothesis for each: https://www.socscistatistics.com/pvalues/tdistribution.aspx. The formulae for calculating t-scores were modified using basic algebra from the formulae for calculating effect sizes from t-scores that Lakens (2013) provides.

Independence of all p-values

Only one p-value was taken from each sample included in each meta-analysis, because each p-value entered into a z-curve must be independent, and if two p-values come from the same sample, they are not independent. If multiple effect sizes were included in a meta-analysis from a single sample, then the first one reported in the table was included in the z-curve and all others taken from that sample were excluded. This means that the number of p-values extracted from each meta-analysis may not equal the number of overall effect sizes included in the original meta-analysis.

Papers that reported multiple meta-analyses

If there were multiple meta-analyses reported in a meta-analysis paper, then separate z-curves were done on each meta-analysis for which the number of effect sizes was 20 or more. Note, however, that, as indicated in the Results below, z-curve will only run if at least 10 of those p-values are below .05.

Disclosure table

There is a disclosure table inspired by those that Simonsohn et al. (2014) recommends for p-curves posted on the OSF. This disclosure table includes information about each meta-analysis and justification for some of the coding decisions. Included in this disclosure table is: (1) the number of z-curves that each paper resulted in; (2) how effect sizes were converted to p-values, with quoted text from each paper to justify it; (3) what subfield the paper was from; (4) the number of times the meta-analysis has been cited on PsycINFO; (5) whether or not the authors tested for publication bias; (6) which tests for publication bias were used; (7) how many different tests of publication bias the authors used; (8) the outcome of each test of publication bias; and (9) the average sample size per study included. There is a separate line of code for each meta-analysis on which a z-curve was performed.

Analytic Technique

Z-curve 2.0 was used for the main analysis (Brunner & Schimmack, 2020). This analysis takes the p-values from a set of studies of interest to the researcher, converts them to two-tailed z-statistics, and uses those z-statistics to calculate average power of all of the studies that have hypothetically been done using finite mixture modeling. Recall

that power is the probability of finding a statistically significant result if there is a true effect (Cohen, 1992), so once the average power is estimated, the percent of all of the studies done in the area that would be predicted to be statistically significant can be estimated. The estimate that z-curve generates, however, is based on the full population of studies predicted to exist, including hypothetical unpublished studies that were not entered in the z-curve. The percent of all (hypothetical) studies predicted to be significant can then be compared to the percent of studies that show statistical significance in the published literature originally entered into the z-curve. If the number of statistically significant results in the published literature is larger than the predicted number of statistically significant results based on the average power of the studies, then this suggests the presence of biases in data-generating mechanisms in the published literature.

In addition to the expected discovery rate (EDR)—the percent of studies predicted to be significant based on the average power of published studies—and the observed discovery rate (ODR)—the percent of studies that show statistical significance in the published literature—z-curve produces several other estimates of interest. But one should note that all of the estimates listed hereafter are simply transformations of the EDR. That is, while they are all interpreted differently, they all basically provide the same information. With that said, the other estimates that z-curve provides are: the expected replication rate (ERR)—the average power of the studies entered, which is also an estimate of the percent of the studies that one would expect to replicate if one performed the studies in exactly the same way as they were done before; the Sorić false discovery rate (Sorić FDR)— the maximum percent of studies that could be false positives; and the file drawer ratio (FDR)—the ratio between the EDR and ODR that is expressed as the number of unpublished findings that are predicted to exist for every published effect size. The FDR is basically a snapshot of the presence of publication bias. For the purposes of reporting the results of this study, the FDR for each z-curve was converted to a percentage. This is done by dividing the FDR by the FDR plus one and then multiplying by 100. This makes interpreting the FDR more straightforward, as the percentage can be interpreted as the percent of findings missing from the published literature. Finally, the value of the EDR subtracted from the ODR was also examined, as the difference between the two is one of the crucial estimates when interpreting a z-curve. The higher this value is, the higher the likelihood of biases in data-generating mechanisms. The z-curves were calculated using the R code uploaded to the OSF (https://osf.io/gr3ax/?view_only=49a2ad228222477abe996cc627ce69b5).

Results

Twenty-nine z-curves were attempted, and 25 of them ran successfully. Four of them did not run because there were fewer than 10 p-values below .05 in the data. The analysis in R will only run if there are at least 10 p-values below .05 in the data entered. Table 1 shows the detailed results for the 25 z-curves that ran successfully. All of the values—except the Significant *N*—are expressed as percentages, and include the 95% confidence intervals. The Significant *N* is included as a column in Table 1 to give an idea of how many p-values went in to each z-curve, as z-curves are run only with the significant p-values that are entered. Note that many of the confidence intervals for the estimates are large, so all interpretations should be cautious.

Out of the 25 z-curves, 19 (76%) showed an ODR larger than the EDR, as indicated by a positive difference between the two. Sixteen (64%) of the z-curves showed a difference greater than 10 percentage points; 11 (44%) showed a difference greater than 20 percentage points; 7 (28%) showed a difference greater than 30 percentage points; and 3 (12%) showed a difference of 50 or more percentage points. Figure 1 shows the values of the ODR – EDR arranged from the highest (70.1) to the lowest (-29.9) values. A negative difference indicates that the power of the published studies was such that one would predict fewer published studies than are available. The more positive the difference, the greater the estimates suggest the presence of biases in data-generating mechanisms.

Figure 2 shows the Sorić FDR for each z-curve, again arranged from lowest (0) to highest (83.8). Thirteen (52%) meta-analyses showed a Sorić FDR below 10; 5 (20%) showed a Sorić FDR between 10 and 20; and 7 (28%) showed a Sorić FDR above 20. The higher the Sorić FDR, the greater the proportion of findings that are predicted to be false positives.

Figure 4 shows the ERR for each z-curve, arranged from lowest (12.4) to highest (100). Eight (32%) of the z-curves showed an ERR below 50%; 13 (52%) showed an ERR below 60%; 19 (76%) showed an ERR below 70%; 21 (84%) showed an ERR below 80%; 23 (92%) showed an ERR below 90%; and 2 (8%) showed an ERR above 90%. The lower the ERR, the lower the percent of studies are predicted to replicate if they were conducted in exactly the same way as they were done the first time.

P-Curves on Three Weakest Meta-Analyses

Because z-curve is a new analytic technique—certainly relative to the older and more widely used p-curve analysis—some may claim that its efficacy and trustworthiness has not yet been demonstrated sufficiently. Therefore, p-curve analysis was also used to examine the distribution of p-values from the three meta-analyses that z-curve analysis indicated to be characterized by the most biases in data-generating mechanisms (i.e., Fischer et al., 2011; Tannenbaum et al., 2015; Weingarten et al., 2016). The metric for determining the three weakest meta-analyses was the FDR, because this captures both QRPs and publication bias. It was decided only to include the three meta-analyses that could be analyzed in full; in other words, no meta-analyses were p-curved if they had multiple meta-analyses reported.

Table 1. Detailed Results for Each Z-curve

Meta-Analysis	Sig. N	ERR	EDR	ODR	SFDR	FDR	ODR – EDR
Balliet et al. (2011) – Punishment	67	69.50[54.90,84.30]	41.70[12.20,77.10]	68.00[58.00, 77.00]	7.30[1.60,3.79]	58.26[22.84,87.80]	26.3[-0.10,45.80]
Balliet et al. (2011) – Reward	13	69.50[38.80,99.80]	40.90[5.00,98.70]	54.00[33.00,74.00]	7.60[0.10,100.00]	59.15[.10,95.00]	13.10[-24.70,28.00]
Brunmair et al. (2019)	96	64.10[48.10,80.20]	48.70[11.20,74.90]	63.00[55.00,70.00]	5.50[1.80,41.70]	51.24[25.09,88.80]	14.30[-4.90,43.80]
Coles et al. (2019)	50	53.10[32.80,68.60]	41.80[16.00,65.40]	36.00[28.00,45.00]	7.30[2.80,27.60]	58.21[34.60,83.99]	-5.80[-20.40,12.00]
Cracco et al. (2018)	171	82.30[73.00,93.80]	75.50[59.60,94.50]	86.00[80.00,90.00]	1.70[0.30,3.60]	24.47[5.48,40.44]	10.50[-4.50,20.40]
Dargue et al. (2019)	46	50.30[29.70,70.50]	32.80[5.00,67.20]	55.00[44.00,66.00]	10.80[2.60,100.00]	67.23[32.84,95.00]	22.20[-1.20,39.00]
Fernandes & Garcia-Marques (2020)	98	65.70[51.30,83.10]	60.30[18.40,80.60]	78.00[70.00,85.00]	3.50[1.30,23.30]	39.65[19.35,81.59]	17.70[4.40,51.60]
Fischer et al. (2011)	25	63.60[32.10,90.70]	12.70[5.00,70.40]	48.00[34.00,62.00]	36.10[2.20,100.00]	87.28[29.58,95.00]	35.30[-8.4,29.00]
Fox et al. (2011) – Performance	20	53.70[19.40,88.10]	52.90[5.00,87.90]	23.00[15.00,33.00]	4.70[0.70,100.00]	47.09[12.05,95.00]	29.90[-54.90,10.00]
Fox et al. (2011) – Solution Time	10	39.40[2.50,87.80]	7.80[5.00,80.70]	38.00[21.00,59.00]	62.20[1.30,100.00]	92.20[19.29,95.00]	30.20[-21.70,16.00]
Hagger et al. (2010)	154	38.70[24.00,53.00]	27.10[8.00,48.10]	78.00[71.00,83.00]	14.10[5.70,60.40]	72.86[51.88,91.98]	50.90[34.90,63.00]
Hu et al. (2020)	42	53.40[25.90,84.10]	48.40[5.00,82.20]	40.00[30.00,50.00]	5.60[1.10,100.00]	51.57[17.83,95.00]	-8.40[-32.20,25.00]
Johnsen & Friborg (2015)	62	90.00[76.60,100.00]	76.10[22.10,100.00]	89.00[78.00,95.00]	1.70[0.00,18.60]	23.90[0.00,77.90]	12.90[-5.00,55.90]
Joseph et al. (2020)	706	83.60[77.10,90.90]	54.00[41.50,83.90]	81.00[78.00,83.00]	4.50[1.00,7.40]	46.00[16.11,58.47]	27.00[-0.90,36.50]
Landau et al. (2015)	36	39.70[14.20,69.00]	13.60[5.00,59.90]	65.00[51.00,77.00]	33.40[3.50,100.00]	86.39[40.12,95.00]	51.40[17.10,46.00]
Lim & Dinges (2010) – Accuracy	15	49.90[13.50,96.70]	46.80[5.00,96.60]	28.00[17.00,43.00]	6.00[0.20,100.00]	53.16[3.38,95.00]	-18.80[-53.60,12.00]
Lim & Dinges (2010) – Speed	13	35.40[5.90,82.80]	23.50[5.00,75.90]	23.00[13.00,37.00]	17.10[1.70,100.00]	76.49[24.07,95.00]	-0.50[-38.90,8.00]
Pan & Rickard (2018)	72	73.50[53.70,91.50]	20.20[6.30,90.80]	52.00[43.00,60.00]	20.80[0.50,78.40]	79.78[9.26,93.71]	31.80[-30.80,36.70]
Pool et al. (2016)	82	49.70[30.60,68.90]	29.50[5.00,64.20]	34.00[28.00,40.00]	12.60[2.90,100.00]	70.54[35.82,95.00]	4.50[-24.20,23.00]
Sedlmeier et al. (2012)	42	62.30[41.70,83.70]	39.60[5.00,80.10]	47.00[37.00,58.00]	8.00[1.30,100.00]	60.41[19.94,95.00]	7.40[-22.10,32.00]
Tannenbaum et al. (2015)	96	53.60[35.20,70.40]	11.00[5.00,35.00]	41.00[34.00,47.00]	42.50[9.80,100.00]	88.98[64.96,95.00]	30.00[12.00,29.00]
Webb et al. (2012) - Experiential	51	35.20[15.80,57.80]	16.80[5.00,54.20]	26.00[20.00,33.00]	26.10[4.50,100.00]	83.22[45.83,95.00]	9.20[-21.20,15.00]
Webb et al. (2012) - Physiological	24	75.10[46.50,95.00]	27.50[5.30,95.90]	62.00[45.00,76.00]	13.90[0.20,94.70]	72.45[4.03,94.74]	34.50[-19.90,39.70]
Weingarten et al. (2016)	100	12.40[2.50,25.10]	5.90[5.00,21.70]	76.00[67.00,83.00]	83.80[19.00,100.00]	94.09[78.31,95.00]	70.10[61.30,62.00]
Zell et al. (2020)	253	100.00[92.00,100.00]	100.00[53.50,100.00]	87.00[82.00,90.00]	0.00[0.00,4.60]	0.00[0.00,46.44]	-13.00[-10.00,28.50]

Note. Sig. *N* = number of p-values entered into the z-curve that were below .05, ERR = expected replication rate, EDR = expected discovery rate, ODR = observed discovery rate, SFDS = Sorić false discover risk, FDR = file drawer ratio, ODR – EDR = expected discovery rate subtracted from the observed discovery rate.

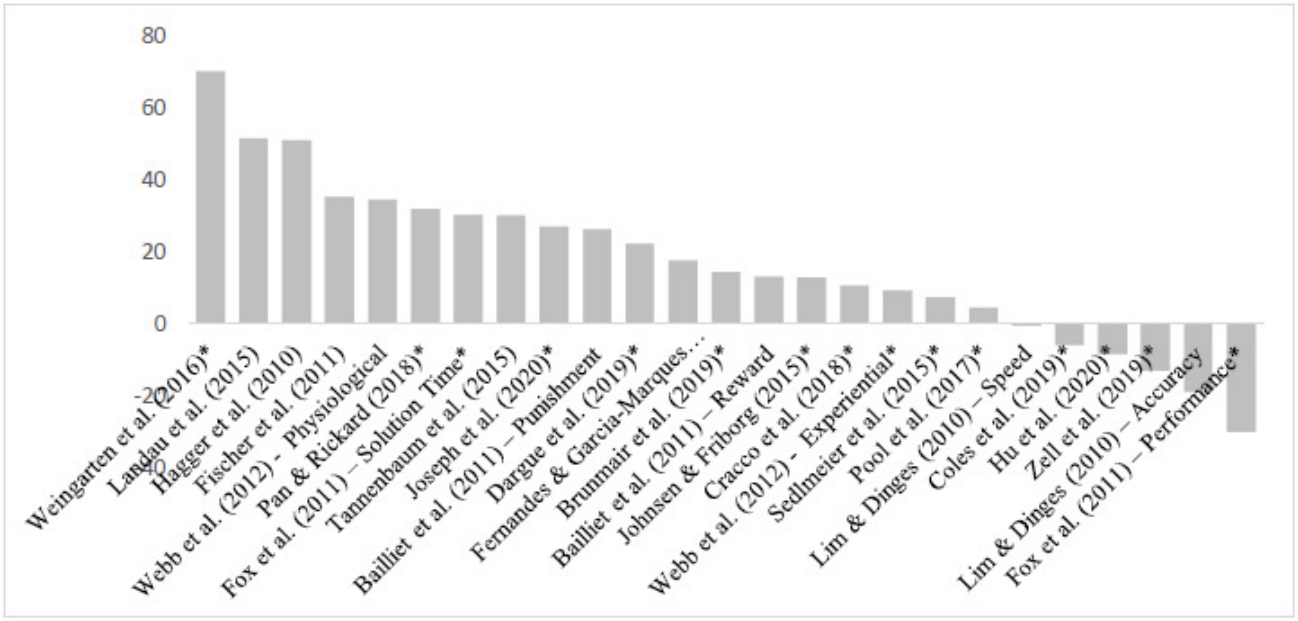

Figure 1. Expected Discovery Rate Subtracted from Observed Discovery Rate for Each Z-curve

The p-curve analysis results for Weingarten et al. appear to corroborate the result of the z-curve. While the half p-curve indicates evidential value, Z = -2.89, p = .002, the 33% power test for the full p-curve is significant, Z = -3.82, p < .001. This suggests that the studies included show less than 33% power, meaning that the studies included are not expected to replicate. Indeed, the p-curve estimates the power for the included studies at 10% [CI: 5%, 18%], which is close to the ERR of 12.4% that the z-curve predicted. Recall that the ERR is essentially the estimated power of the studies included in a z-curve, and so it should be equivalent to the average power estimated by p-curve.

The results for Tannenbaum et al. and Fischer et al. are more complicated. The full p-curve, Z = -18.24, p < .001, and the half p-curve, Z = -20.66, p < .001, are both significant for Tannenbaum et al.. The 33% power test also came out favorably for Tannenbaum et al. in both its full p-curve, Z = 10.75, p > .999, and its half p-curve, Z = 19.66, p > .999. Its estimated power is 91% [87%, 94%]. The p-curve results for Fischer et al. were similarly favorable. The full p-curve, Z = -7.51, p < .001, and half p-curve, Z = -9.26, p < .001, were both significant, and the 33% power test also came out favorably for both the full p-curve, Z = 3.99, p > .999, and the half p-curve, Z =8.92 , p > .999. The estimated power for Fischer et al. was 84% [68%, 93%]. The discrepancy in the outcomes of the z-curves and the outcomes of the p-curves is likely because the results of p-curves can be problematic when there is a lot of heterogeneity in the effect sizes in the population of studies being analyzed, as mentioned in the Introduction (Brunner & Schimmack, 2020). P-curve's estimates can also be upwardly biased when there are a few highly-powered studies entered (Brunner & Schimmack, 2020). Thus, the discrepancy in the outcomes should not be taken as too strong a reason to question the outcomes of the z-curves reported above.

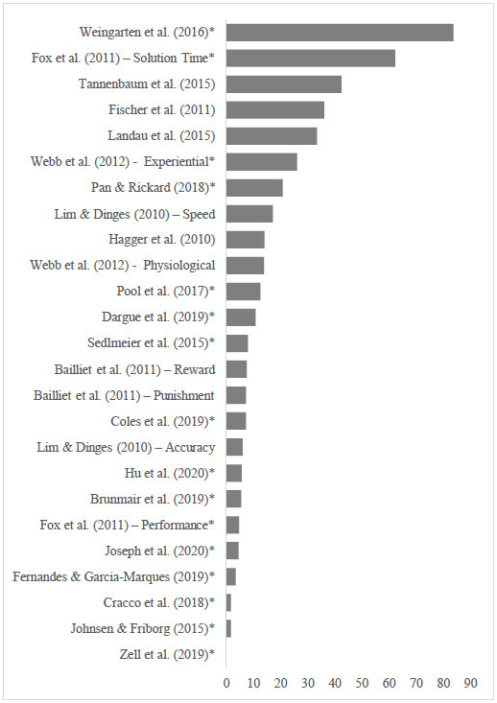

Figure 2. Sorić False Discovery Risk for Each Z-curve

Discussion

The z-curve analyses of the studies included in 21 influential meta-analytic reviews suggest the presence of questionable research practices (QRPs) in a majority of the examined literatures, and that a non-trivial proportion of studies included are predicted not to replicate. Three-quarters of the z-curves suggest that publication bias has resulted in one unpublished study for every published study, suggesting that only half of all studies have been published

(i.e., FDR of 50% or higher). Around one-third of the z-curves suggest that up to 50% of the studies included are not expected to replicate as indicated by an ERR below 50%, and a similar proportion of the z-curves suggest that over 20% of studies included may be false positives. In aggregate, these results suggest a relatively widespread prevalence of biases in the data-generating mechanisms of the included primary studies—and hence suggest a similar bias in meta-analytic effect size estimates.

While the results show that QRPs appear to be quite widespread, they also appear to vary across literatures. Some meta-analyses were characterized by excellent estimates of the expected replication rate and low file drawer ratios such that the meta-analytic effect size estimates are likely to be relatively unbiased. The results of z-curve analyses for other meta-analyses were more concerning, suggesting low replicability and high file drawer ratios. While these data cannot speak to the reasons for this high variability in the results of the z-curve analyses, some possible reasons present themselves.

First, it may be that some literatures focus on effects that are inherently larger and more robust. In such settings, there may be no need to engage in QRPs. When studying a large effect, a researcher should not need QRPs to find an effect in their study (Cohen, 1992). Second, research and publication norms are likely to vary across subdisciplines. That is, in some subdisciplines, researchers may believe that QRPs are acceptable and widely practiced, and that journals have a strong preference for statistically significant findings. In others, researchers may believe that journals are more accepting of null results, or there may be strong research norms in place that prevent QRPs. Third, it may be that data collection is more time consuming, difficult, and expensive in some subdisciplines, which may make certain practices such as "data peeking" and optional stopping rules seem more acceptable. Researchers in these subdisciplines may not want to use up any more time, money, or resources than are totally essential, and so it may seem to make sense to check for statistical significance periodically and stop once it is achieved to avoid being wasteful by collecting more data than necessary.

The results of two of the three p-curves reported above were much more favorable than their corresponding z-curves. This is likely because an assumption of p-curve is that the studies included all investigate a single underlying effect; in other words, p-curve assumes effect size homogeneity. Brunner & Schimmack (2020) demonstrate that p-curves can overestimate evidential value when this assumption is not met—in other words, when there is effect size heterogeneity. Indeed, p-curve can overestimate statistical power by almost 25% under certain levels of effect size heterogeneity, whereas z-curve performs well under conditions of heterogeneity (Brunner & Schimmack, 2020). Every single meta-analysis coded for this work reported estimates of effect size heterogeneity, and all indicated the presence of heterogeneity. Thus, it stands to reason that p-curve may not be the ideal way to test the evidentiary value of those studies.

This work is not meant to encourage anyone to discount all published meta-analyses. Neither is it meant to point to any inherent problems with meta-analysis as an approach.

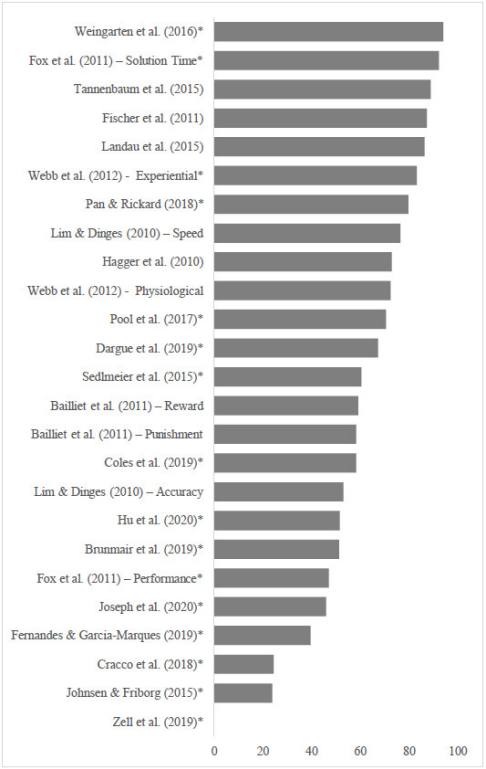

Figure 3. File Drawer Ratios for Each Z-curve

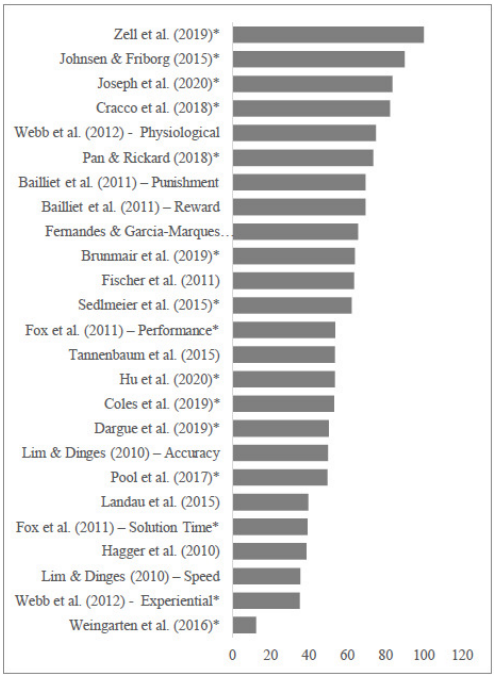

Figure 4. Expected Replication Rate for Each Z-curve

It is merely to show that biases in the data-generating mechanisms of primary studies in some literatures appear to have affected the meta-analytic syntheses of those literatures. This, in turn, suggests that these meta-analytic results should be interpreted with some caution. In particular, it seems likely that effect size estimates are inflated, which

has both practical implications (e.g., should an expensive intervention with a smaller effect size still be conducted) and methodological implications inasmuch as researchers may be basing their power calculations on inflated effect size estimates.

Contributions

The first author came up with the idea for the project, performed all of the coding and analyses, and wrote the manuscript.

Acknowledgments

Grateful acknowledgment is extended to Marcus Credé for providing invaluable feedback on drafts of this manuscript.

Competing Interests

The author has no competing interests to state.

Data Accessibility

All data analyzed for this project, as well as the R code used for the analyses, are available on the Open Science Framework at the following link: https://osf.io/gr3ax/?view_only=49a2ad228222477abe996cc627ce69b5.

Submitted: October 26, 2021 PST, Accepted: February 08, 2022 PST

REFERENCES

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*. https://doi.org/10.15626/mp.2018.874

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A mea-meta-analysis: Empirical review of statistical power, Type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*(2), 239–270. https://doi.org/10.1080/00273171003680187

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chevret, S., Ferguson, N. D., & Bellomo, R. (2018). Are systematic reviews and meta-analyses still useful research? No. *Intensive Care Medicine*, *44*(4), 515–517. https://doi.org/10.1007/s00134-018-5066-3

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, *145*(6), 610–651. https://doi.org/10.1037/bul0000194

Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., Radkova, I., Deschrijver, E., & Brass, M. (2018). Automatic imitation: A meta-analysis. *Psychological Bulletin*, *144*(5), 453–500. https://doi.org/10.1037/bul0000143

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, *145*(8), 765–784. https://doi.org/10.1037/bul0000202

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*(5), 517–517. https://doi.org/10.1037/0003-066x.33.5.517.a

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. https://doi.org/10.1037/a0024445

Fernandes, A. C., & Garcia-Marques, T. (2020). A meta-analytical review of the familiarity temporal effect: Testing assumptions of the attentional and the fluency-attributional accounts. *Psychological Bulletin*, *146*(3), 187–217. https://doi.org/10.1037/bul0000222

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517–537. https://doi.org/10.1037/a0023304

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.3102/0013189x005010003

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525. https://doi.org/10.1037/a0019486

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Hu, X., Cheng, L. Y., Chiu, M. H., & Paller, K. A. (2020). Promoting memory consolidation during sleep: A meta-analysis of targeted memory reactivation. *Psychological Bulletin*, *146*(3), 218–244. https://doi.org/10.1037/bul0000223

Ioannidis, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

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

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin*, *141*(4), 747–768. https://doi.org/10.1037/bul0000015

Joseph, D. L., Chan, M. Y., Heintzelman, S. J., Tay, L., Diener, E., & Scotney, V. S. (2020). The manipulation of affect: A meta-analysis of affect induction procedures. *Psychological Bulletin*, *146*(4), 355–375. https://doi.org/10.1037/bul0000224

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., Hasselman, F., Corker, K. S., Grange, J., Sharples, A., Cavender, C., Augusteijn, H., Augusteijn, H., Gerger, H., Locher, C., Miller, I. D., Anvari, F., & Scheel, A. M. (2017). *Examining the reproducibility of meta-analyses in psychology: A preliminary report*. https://doi.org/10.31222/osf.io/xfbjf

Landau, M. J., Kay, A. C., & Whitson, J. A. (2015). Compensatory control and the appeal of a structured world. *Psychological Bulletin*, *141*(3), 694–722. https://doi.org/10.1037/a0038703

Lim, J., & Dinges, D. F. (2010). A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological Bulletin*, *136*(3), 375–389. https://doi.org/10.1037/a0018883

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Olsson-Collentine, A., van Aert, R. C. M., Bakker, M., & Wicherts, J. M. (2021). *Preprint - Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting*. https://doi.org/10.31234/osf.io/43yae

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157. https://doi.org/10.2307/1164923

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin*, *142*(1), 79–106. https://doi.org/10.1037/bul0000026

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: A meta-analysis. *Psychological Bulletin*, *138*(6), 1139–1171. https://doi.org/10.1037/a0028168

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Sharpe, D., & Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 377–387. https://doi.org/10.1037/cap0000215

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

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/10.1037/a0033242

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). John Wiley & Sons Ltd.

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, *141*(6), 1178–1204. https://doi.org/10.1037/a0039729

Webb, T. L., Miles, E., & Sheeran, P. (2012). Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychological Bulletin*, *138*(4), 775–808. https://doi.org/10.1037/a0027600

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, *142*(5), 472–497. https://doi.org/10.1037/bul0000030

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, *146*(2), 118–149. https://doi.org/10.1037/bul0000218

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/32571-garbage-in-garbage-out-evaluating-the-evidentiary-value-of-published-meta-analyses-using-z-curve-analysis/attachment/82122.docx?auth_token=9ulffiZhL8pInduyFAy8