Investigating the Reproducibility of Psychological Meta-Analyses

E. Maassen + Truth squad

31-01-19

The ever-increasing growth of scientific publication output (Bornmann, 2015) has led to an overload and need for systematic reviewing of evidence. Meta-analysis is a widely used method to synthesize quantitative evidence from multiple individual studies. Meta-analysis involves a set of procedural and statistical techniques to arrive at an overall effect size estimate, and can be used to inspect whether study outcomes differ systematically based on particular study characteristics (Hedges & Olkin, 1985). However, careful considerations are needed when conducting a meta-analysis, because of the many (sometimes relatively arbitrary) decisions and judgments that one has to make during various stages of the research process. Procedural differences in coding primary studies could lead to variation in results, which affects the validity of drawn conclusions (Valentine, Cooper, Patall, Tyson, & Robinson, 2010). Likewise, meta-analysts often need to perform complex computations to synthesize primary study results to get an overall meta-analytic outcome, which creates the risk of faulty data handling and erroneous estimates. When decisions and calculation steps are not carefully undertaken and reported on in detail, the methodological quality of the meta-analysis cannot be assessed, and reproducibility is undermined by possible occurring biases arising from inconsistent decisions, inaccurate recalculations and reporting errors. Our aim was to find out how reproducible primary study effect size estimates in psychological meta-analytic articles are, and whether irreproducibility of these primary study effect sizes affects meta-analytic outcomes.

Research from various fields has demonstrated issues arising from substandard reporting practices: in biomedical research the transformations from primary study effect sizes to the standardized mean difference (SMD) used in meta-analyses were often inaccurate (Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007). This meant that effect size estimates often could not be reproduced, which also had an effect on the pooled meta-analytic effect size. Within organizational sciences, evidence is found for incomplete and nontransparent meta-analytic reporting practices (Aytug, Rothstein, Zhou, & Kern, 2011; Geyskens, Krishnan, Steenkamp, & Cunha, 2008). Even though the numerous choices and judgment calls meta-analysts make do not always influence substantive conclusions (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2010), it is still important to investigate the extent and source of irreproducibility of primary study effect sizes. Within psychology, results appear to be similar to other fields of research: a severe lack of relevant information related to effect size extraction, coding, and adhering to reporting guidelines hinders and sometimes prohibits any meta-analytic effect size reproducibility (Lakens et al., 2017).

This project builds upon previous efforts examining effect size computational errors and reproducibility of meta-analyses. In the first study we considered effect sizes reported in meta-analytic articles and searched for the primary study articles to check whether we could recompute the effect size as reported in the meta-analytic articles; this we call primary study effect size reproducibility. In the second study we considered if corrections in these primary study effect sizes affected main meta-analytic outcomes, namely the effect size estimate, confidence intervals, and heterogeneity estimates. We call this meta-analysis reproducibility. To this end, we checked the availability of necessary information in the primary study article to calculate primary study effect sizes, we transformed these primary study effect sizes to a standardized effect size estimate, and we reran meta-analyses to see if meta-analytic outcomes were reproducible based on our recalculated results.

Study 1. Primary study effect size reproducibility.

In study 1 we documented primary study effect sizes as they are reported in meta-analyses (i.e., in a data table) and attempted to reproduce them based on the information given in the meta-analysis. There exist several reasons for limited primary study effect size reproducibility. First, the primary study article may lack

sufficient information to reproduce the effect size (e.g., missing standard deviations). Second, it might be unclear which steps the meta-analysts conducted. It could be the case the primary paper contains multiple relevant effect sizes but it is unclear which effect is included in the meta-analytic article due to ambiguous reporting, either in the demarcation of inclusion criteria or imprecision when stating which specific effect was extracted from the primary study paper. Oppositely, it could be clear which effect from the primary paper is included in the meta-analysis, but a given primary study can yield different effect size computations or transformations standardized effect size, and it is unclear what calculations were made by the meta-analysts. Third, potentially a calculation error was made when computing the primary study effect size for inclusion in the meta-analytic article.

Our hypotheses, design, and analysis plan were preregistered and can be found at https://osf.io/v2m9j. We hypothesized that a sizeable proportion of reproduced primary effect sizes would be different from the original calculation of the authors because of ambiguous reporting or potential errors in effect size transformations (Gøtzsche et al., 2007). We also expected more discrepancies in primary studies that used effect size estimates that require more calculation steps (i.e., SMDs) compared to effect sizes that are often extracted as is (i.e., correlations), as well as more potential errors in unpublished primary studies compared to published primary studies because the former contain varying reporting standards and are sometimes not peer-reviewed. Our goal was to figure out what percentage of the primary study effect sizes within meta-analytic articles is reproducible, and what causes primary study effect size irreproducibility.

Method

Sample

Meta-analysis selection

Our sample of meta-analyses (total $m = 33$)[1] was chosen from previous research that had coded materials readily available (Bakker, van Dijk, & Wicherts, 2012; Wijsen, 2015). The goal of the meta-analysis selection was to obtain a representative sample of psychological studies, but our inclusion criteria may have affected the representativeness of our sample. More specifically, a large number ($m = 97$) of meta-analyses could not be included because they omitted necessary basic statistics, such as a data table with primary studies or appropriate effect sizes. In total, 23 of the included meta-analyses were selected by Wijsen (2015), whereas the final 10 meta-analyses were selected from previous research by Bakker et al. (2012). The total number of selected meta-analytic articles was 33, which contained 1951 primary effect sizes, of which we sampled 500 to reproduce.

Primary study selection

When selecting primary studies from the meta-analyses, we wanted to sample both primary studies that could be considered outliers compared to the rest of the primary studies in the meta-analysis, as well as non-outlier primary studies. Some data errors in meta-analyses are likely to result in outliers, which inflate variance estimates (Schmidt & Hunter, 2015). Because of this, we oversampled outlier primary studies to investigate whether outliers were more likely to be erroneous compared to non-outlier primary studies[2]. For all meta-analyses, we firt classified all primary studies as either being an outlier or non-outlier. We then randomly sampled from the collection of outlier primary studies per meta-analysis. To ensure a fair distribution of outlier and non-outlier primary studies, we sampled 10 outlier primary studies per meta-analysis if that many could be obtained. In total, we included 197 primary studies that were considered outliers. After this selection, a total of 303 primary studies remained to be sampled from the non-outlier

[1] A detailed sampling scheme and flowcharts can be found at https://github.com/emaassen/paper-effectsizes/blob/master/appendices/a-sampling-scheme.pdf.

[2] Note that our primary study sample is not completely random because we first divided our sample into two strata (outlier and non-outlier), from which we took random samples.

primary studies within the 33 meta-analyses. The remaining primary studies were evenly selected from the meta-analyses, meaning we randomly sampled approximately 10 non-outlier primary study effect sizes per meta-analysis.

Procedure

We set out to reproduce 500 primary study effect sizes from 33 meta-analyses. Our procedure of recalculating primary study effect sizes is displayed in Figure 3. In all cases, we strictly followed methods as described in the meta-analytic article to recompute the primary study effect sizes. In cases where the meta-analysis was unclear as to which methods were used, we employed well-known formulas for transformation[3]. We assumed group sizes were equal if no further information was given in the meta-analysis. We adjusted all effects in such a way that insofar the prediction of the meta-analysts was corroborated, the mean effect size would be positive, which entailed reversing the effect sizes of five meta-analyses. Consequently, a negative effect size indicates a different direction of effect than was expected.

Our primary outcome of interest was the discrepancy between the reported and reproduced effect size estimates. We classified the discrepancies between original and reproduced effect sizes as small, moderate, and large. Recent meta-analyses in various psychological fields (i.e., consumer, educational, health, personality and social psychology) show Hedges' g effect sizes commonly range from .10 to .50 (e.g., Mitchell, Amaro, & Steele, 2016; Shariff, Willard, Andersen, & Norenzayan, 2015; Spangenberg, Kareklas, Devezer, & Sprott, 2016; Wanzek et al., 2015). Based on these results, we chose the classifications of discrepancies in Hedges' g to be small [.050 - .149], moderate [.150 - .249] and large [.250 - ∞] and transformed them to similar classifications for the other effect sizes[4].

We categorized reproduced primary study effects into one of four categories: (0) we could reproduce the effect size as reported in the meta-analytic article; (1) not enough information was available to reproduce the effect size (e.g., *SD*s are missing in the primary article). In this case we copied the original effect size as reported in the meta-analysis because we were not sure which computations the meta-analysts performed, or whether they had contacted the authors for necessary statistics; (2) it was unclear which steps the meta-analysts conducted (e.g., multiple appropriate effects were found in the primary study, and it was unclear which one was extracted). When this occurred we inferred which primary study effects would be pertinent for inclusion in the meta-analysis, and chose the primary study effect size that most resembled the effect size as reported in the meta-analysis; (3) our recalculation resulted in a different effect size (i.e., a potential calculation error was made). Two coders independently attempted to reproduce the 500 primary study effect sizes. Differences or disagreements in coding were solved after discussion. The interrater reliability for discrepancies in calculated effect sizes was .97. The interrater reliability for discrepancies in which category the effect size belonged to (i.e., a reproduced, different, incomplete or ambiguous effect was) $\kappa = .71$.

[3]The information we extracted and formulas we used to compute and transform all primary study effect sizes can be found for each meta-analysis on https://osf.io/a9dw4/.

[4]Appendix D displays Figure 4 in more detail, separated by the four types of effect sizes found in the meta-analytic articles (i.e., SMDs d and g, correlations r and Fisher's corrected correlation z). The results as illustrated in Figure 4 are reported per meta-analysis in Appendix D.

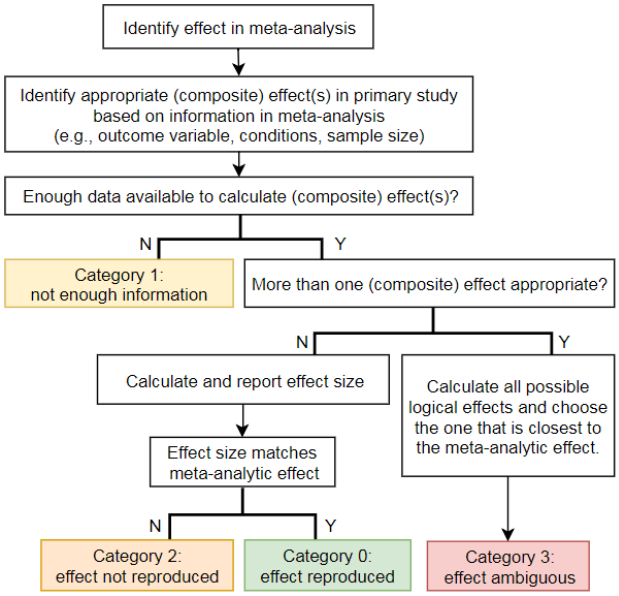

Figure 1: Figure 1. Decision tree of primary study effect size recalculation and classification of discrepancy categories.

Results

Results showed that many primary study effect sizes could not be reproduced. Out of 500 sampled primary study effect sizes, we could reproduce 276 (55%) without any issues, whereas for 224 (45%) effect sizes we could not. Figure 4 displays all primary study effect sizes that used either Cohen's d or Hedges' g as the meta-analytic effect size ($k = 247$), where all primary study effect sizes using Cohen's d were transformed to Hedges' g. Likewise, Figure 5 displays all primary study effect sizes that used either a product-moment correlation r or Fisher's r-*to*-z transformed correlation ($k = 253$), where all primary study effect sizes using a product-moment correlation r were transformed to Fisher's r-*to*-z correlation[5].

For 54 effect sizes (11%), the primary study paper did not contain enough information to reproduce the effect, so the original effect size was copied. For 74 effect sizes (15%), a different effect than originally stated in the meta-analytic article was calculated, while for 96 effect sizes (19%) it was unclear what procedure was followed by the meta-analysts, and the effect size which was most relevant and closest to the effect size was chosen.

The most common reason for not being able to reproduce a primary study effect size was missing or unclear information in the meta-analysis. More specifically, it was often unclear which specific effect was extracted from the primary study because multiple effects were relevant to the research question. Similarly, it was often unclear which sample or time point was included. It was not always clear whether the included effect was constructed from a combination of multiple effects, or how the primary study effect was transformed to the effect size included in the meta-analysis. Other prevalent issues pertaining to potential data errors or lack or clarity in the meta-analytic process were inconsistencies in inclusion criteria within a meta-analytic article, inconsistencies in how sample sizes were calculated, the inclusion of the same sample of respondents for multiple effects without correction, reporting the use of formulas or corrections that were not used, and not reporting the direction of variables. Notably, for 66 primary study effect sizes (13%), the reported and reproduced sample sizes did not coincide. Additionally, we inspected possible sample dependency in our meta-analyses, through checking the samples of pairs of primary studies that were flagged as having

[5]Results for each meta-analysis separately can be found at https://github.com/emaassen/paper-effectsizes/blob/master/appendices/c-individual-scatterplots.pdf.

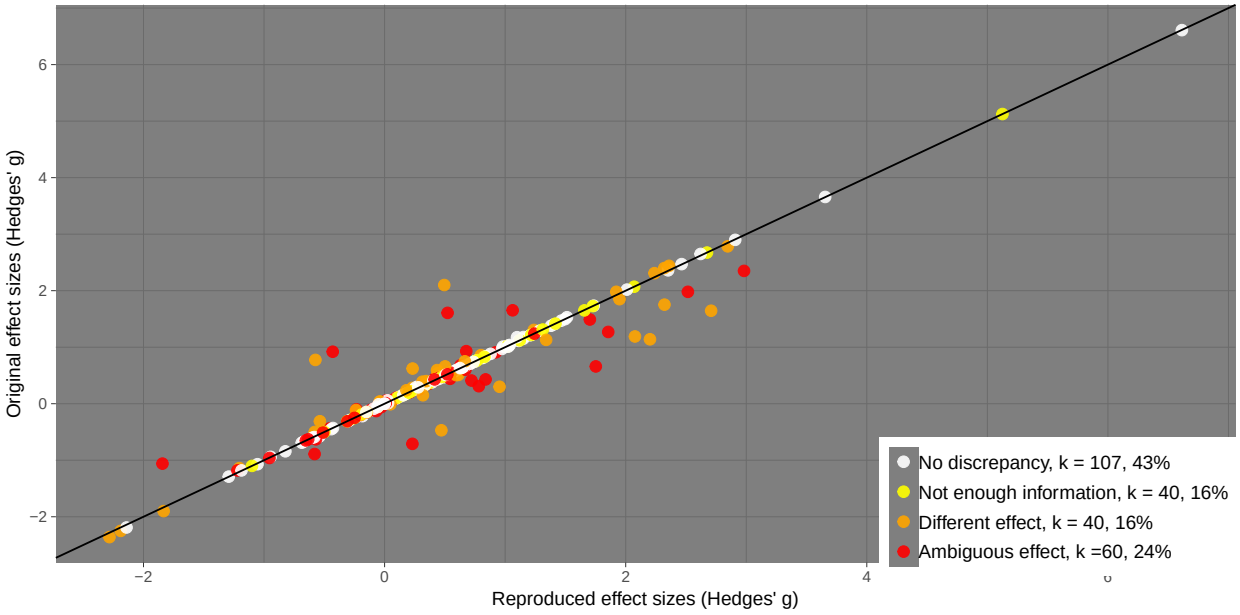

Figure 2: Figure 2. Scatterplot of 247 original and reproduced standardized mean difference effect sizes from 33 meta-analyses. All effect sizes are transformed to Hedges' g.

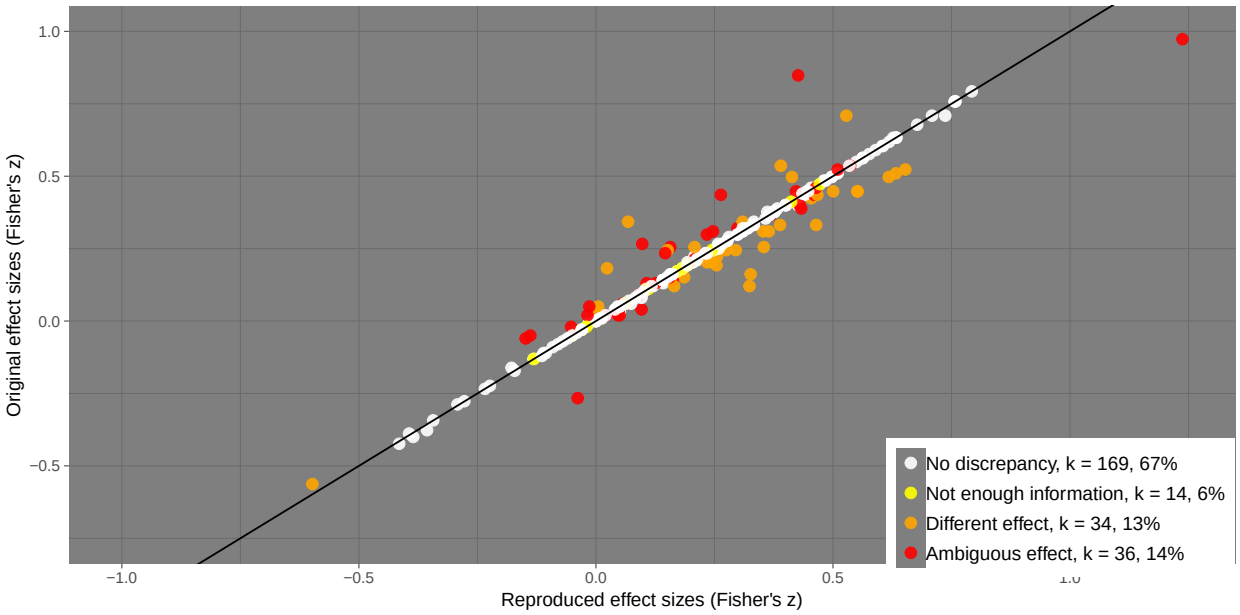

Figure 3: Figure 3. Scatterplot of 253 original and reproduced correlation effect sizes from 33 meta-analyses. All effect sizes are transformed to Fisher's z.

Table 1: Table 1. Descriptive statistics on the reproducibility and characteristics of primary studies.

	Irreproducible		Reproducible		Total	
SMD	140	(57%)	107	(43%)	247	(100%)
correlation	84	(33%)	169	(67%)	253	(100%)
outlier	77	(39%)	120	(61%)	197	(100%)
non-outlier	147	(49%)	156	(51%)	303	(100%)
published	216	(47%)	239	(53%)	455	(100%)
unpublished	8	(18%)	37	(82%)	45	(100%)

homogeneous effect sizes and at least one author in common.[6] Over all meta-analyses combined, we found a total of 248 possible pairs of dependent primary study effect sizes. Of those 248 pairs, we were unable to locate 47, leaving 201 pairs to be investigated. Of these 201 pairs, 19 were indicated as being likely to have used the same sample of participants for multiple effects. In total, 11 out of 33 meta-analyses contained pairs of primary studies with possible dependent effect sizes.

Table 1 contains descriptive statistics on reproducibility and various primary study characteristics. We expected primary studies with SMDs to be less reproducible than studies with other effect sizes. In total, 59% of all primary studies containing SMDs were irreproducible, whereas for primary studies with correlations this was 30%. We expected less reproducibility in primary studies classified as outliers (compared to non-outliers). However, contrary to our expectation, 35% of all outlier effect sizes was irreproducible, whereas 49% of non-outlier effect sizes was irreproducible[7]. Finally, we expected effect sizes from unpublished studies to be less reproducible compared to published studies, but we found 13% of unpublished studies and 47% of published studies were irreproducible.

In total, 120 out of 500 primary study effect sizes (24%) showed discrepancies compared to the primary study effect sizes as reported in the meta-analytic articles. Of those 120 discrepancies, 49 were small, 14 were moderate and 57 were large. We note that it is possible for a primary study effect size to be classified as irreproducible, even if we found no discrepancies between the reported and recalculated primary study effect size estimate. For instance, this is the case with all primary study effect sizes that did not contain enough information to reproduce (making it irreproducible), thus we had to copy the reported effect size (which means no discrepancies were found).

Because we oversampled outliers, our sample of 500 primary study effect sizes is not representative for the population of 33 meta-analyses we sampled from. In the population, 30% of effect sizes classifies as an outlier, compared to 39% of our sample, meaning our sample contains too many outlier primary study effect sizes, and too few non-outliers. In order to obtain the chance of any given effect size in a certain meta-analysis being irreproducible, we need to deal with this overrepresentation of outliers by design. To adjust for this, we first calculated correction weights using type of effect size (outlier or non-outlier) as the auxiliary variable[8]. We adjusted the sample proportions for each meta-analysis separetely so they were in line with their corresponding population proportions, and used these estimates to calculate the probability of finding a potential error (i.e., either a different, incomplete, or ambiguous effect size) in a primary study in each meta-analysis separately, and over all 33 meta-analyses in total. The computed error probability for the 33 meta-analyses varied from 0 to 1. Across all meta-analyses, the chance of any randomly chosen primary study effect size being erroneous is 0.37.

[6]The full document on sample dependency with further explanation can be found at https://github.com/emaassen/paper-effectsizes/blob/master/appendices/d-dependency-samples.pdf.

[7]Note that we did not preregister any hypotheses on whether (non-)outlier primary study effect sizes were reproducible or not, and this result should be considered an explorative finding.

[8]More detailed information on how we corrected for oversampling can be found at https://github.com/emaassen/paper-effectsizes/blob/master/appendices/e-oversampling.pdf.

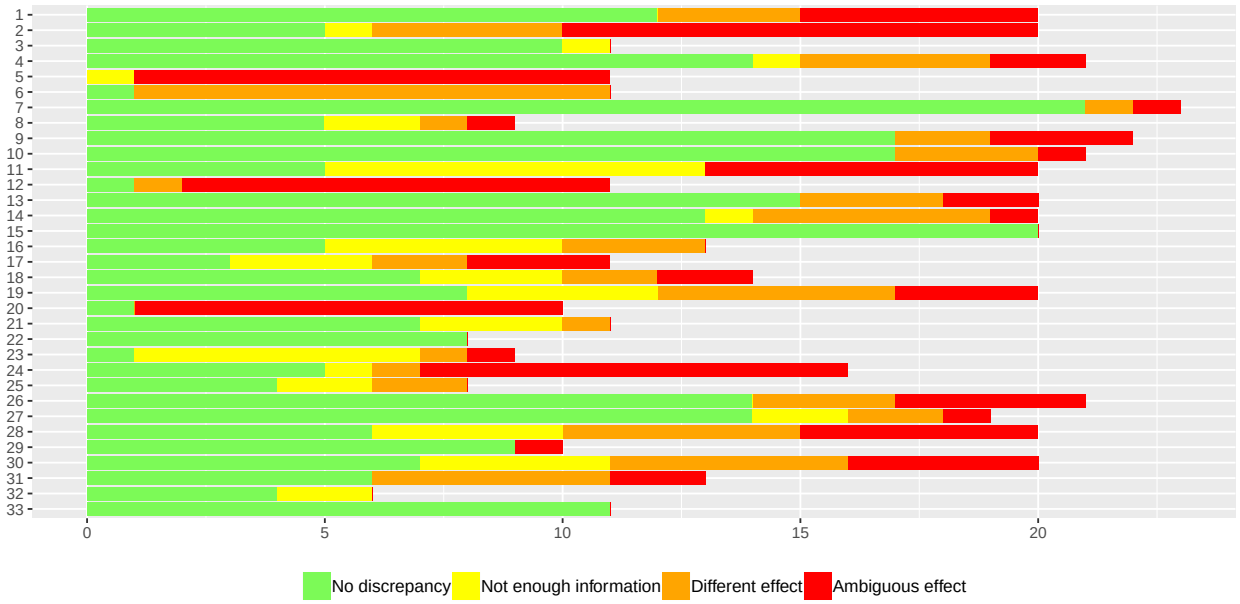

Figure 4 displays a bar plot with the frequency of discrepancies per meta-analysis; only three of the 33 samples of primary studies were completely reproducible, 29 were partly reproducible, and for one, none of the sampled primary study effect sizes was reproducible.

Discussion

We set out to investigate to what extent primary study effect sizes are reproducible. In total, we sampled 500 primary study effect sizes from 33 meta-analyses. Of the 500 recalculated primary study effect sizes, 320 (44%) could not be reproduced. The main reasons were either because necessary statistical data were omitted from the primary study, necessary information regarding the effect size selection or computation was missing from the meta-analysis, or a different effect was found, due to potential extraction, transformation, or reporting errors. We acknowledge that our sample of reproduced primary study effect sizes are not necessarily the primary study effect sizes that were intended by the meta-analytic authors; in many cases the reporting on which specific effect was extracted from the primary study was very vague, and which effects we deemed fitting might differ substantially from the intended effects from the authors. This lack of transparent and clear reporting regarding the primary study effect sizes is concerning, especially when considering many meta-analyses were excluded from our sample due to not reporting their data table. In our sample of relatively well-reported meta-analyses, the chance of a randomly chosen study being irreproducible is 0.37. We can assume in these omitted meta-analyses the reporting standards and irreproducibility probabilities are even worse. Poor reproducibility at the primary study level might affect meta-analytic outcomes, which is worrisome because it could bias the meta-analytic evidence and lead to substantial changes in conclusions. Our goal was to study what the effect of the irreproducibility of primary study effect sizes would be on meta-analytic outcomes, which will be discussed in study 2.

Study 2. Meta-analytic outcomes reproducibility.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2010). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*(1), 5–38. https://doi.org/10.1177/0149206310377113

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2011). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, *15*(1), 103–133. https://doi.org/10.1177/1094428111403495

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

Bornmann, M., L. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Geyskens, I., Krishnan, R., Steenkamp, J.-B. E. M., & Cunha, P. V. (2008). A review and evaluation of meta-analysis practices in management research. *Journal of Management*, *35*(2), 393–419. https://doi.org/10.1177/0149206308328501

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA*, *298*(4). https://doi.org/10.1001/jama.298.4.430

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Lakens, D., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., Hasselman, F., … Scheel, A. M. (2017). Examining the reproducibility of meta-analyses in psychology: A preliminary report. https://doi.org/10.31222/osf.io/xfbjf

Mitchell, T. B., Amaro, C. M., & Steele, R. G. (2016). Pediatric weight management interventions in primary care settings: A meta-analysis. *Health Psychology*, *35*(7), 704–713. https://doi.org/10.1037/hea0000381

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. SAGE Publications, Ltd. https://doi.org/10.4135/9781483398105

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. (2015). Religious priming. *Personality and Social Psychology Review*, *20*(1), 27–48. https://doi.org/10.1177/1088868314568811

Spangenberg, E. R., Kareklas, I., Devezer, B., & Sprott, D. E. (2016). A meta-analytic synthesis of the question-behavior effect. *Journal of Consumer Psychology*, *26*(3), 441–458. https://doi.org/10.1016/j.jcps.2015.12.004

Valentine, J. C., Cooper, H., Patall, E. A., Tyson, D., & Robinson, J. C. (2010). A method for evaluating research syntheses: The quality, conclusions, and consensus of 12 syntheses of the effects of after-school programs. *Research Synthesis Methods*, *1*(1), 20–38. https://doi.org/10.1002/jrsm.3

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2015). Meta-analyses of the effects of tier 2 type reading interventions in grades k-3. *Educational Psychology Review*, *28*(3), 551–576. https://doi.org/10.1007/s10648-015-9321-7

Wijsen, L. (2015). *Do psychological effects fade away? Meta-analyzing the decline effect.*