Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report

Daniel Lakens, Eindhoven University of Technology, NL

Marcel van Assen, Tilburg University & Utrecht University, NL

Farid Anvari, Flinders University, Australia

Katherine S. Corker, Kenyon College, USA

James A. Grange, Keele University, UK

Heike Gerger, Universität Basel, Switzerland

Fred Hasselman, Radboud University Nijmegen, NL

Jacklyn Koyama, University of Toronto, Canada

Cosima Locher, Universität Basel, Switzerland

Ian Miller, University of Toronto, Canada

Elizabeth Page-Gould, University of Toronto, Canada

Felix D. Schönbrodt, Ludwig-Maximilians-Universität München, Germany

Amanda Sharples, University of Toronto, Canada

Barbara A. Spellman, University of Virginia, USA

Shelly Zhou, University of Toronto, Canada

Author Note

Funding for this research was provided by the Berkeley Initiative for Transparency in the Social Sciences, a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation.

All materials are available from https://osf.io/q23ye/

Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, IPO 1.33, PO Box 513, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl.

Study Abstract

Meta-analyses are an important tool to evaluate the literature. It is essential that meta-analyses can easily be reproduced to allow researchers to evaluate the impact of subjective choices on meta-analytic effect sizes, but also to update meta-analyses as new data comes in, or as novel statistical techniques (for example to correct for publication bias) are developed. Research in medicine has revealed meta-analyses often cannot be reproduced. In this project, we examined the reproducibility of meta-analyses in psychology by reproducing twenty published meta-analyses. Reproducing published meta-analyses was surprisingly difficult. 96% of meta-analyses published in 2013-2014 did not adhere to reporting guidelines. A third of these meta-analyses did not contain a table specifying all individual effect sizes. Five of the 20 randomly selected meta-analyses we attempted to reproduce could not be reproduced at all due to lack of access to raw data, no details about the effect sizes extracted from each study, or a lack of information about how effect sizes were coded. In the remaining meta-analyses, differences between the reported and reproduced effect size or sample size were common. We discuss a range of possible improvements, such as more clearly indicating which data were used to calculate an effect size, specifying all individual effect sizes, adding detailed information about equations that are used, and how multiple effect size estimates from the same study are combined, but also sharing raw data retrieved from original authors, or unpublished research reports. This project clearly illustrates there is a lot of room for improvement when it comes to the transparency and reproducibility of published meta-analyses.

Study Description

Meta-analysis has long been thought of as an important tool to evaluate a research literature, and researchers increasingly recognize that psychology needs to move towards meta-analytic thinking (Open Science Collaboration, 2015; Simons, Holcombe, & Spellman, 2014). Although meta-analyses do not provide definitive conclusions, they are often interpreted as state-of-the-art knowledge about a research topic. Meta-analyses are widely cited, assumed to provide the most reliable estimate of the population effect size, and influence theory development. It is therefore important to make sure published meta-analyses are of the highest possible quality.

The recent recognition of how the inherent uncertainty in scientific findings is amplified by undesirable research practices, and the subsequent efforts to improve our empirical knowledge base, have had two contradictory effects on the perceived value of meta-analyses. On one hand, given the broad array of problems that make it difficult to evaluate the evidential value of single studies, it seems more imperative than ever to use meta-analytic techniques to evaluate larger bodies of research if we want to get an accurate view of the size of an effect, its robustness, and possible moderators. On the other hand, some researchers have doubted whether meta-analyses can actually produce objective knowledge when the selection, inclusion, coding, and interpretation of studies leaves as least as much room for flexibility in the analysis and conclusions about the results of the meta-analysis as is present in single studies (e.g., Ferguson, 2014). Ioannidis (2016) has argued most meta-analyses and systematic reviews are flawed, redundant, or misleading.

A lack of openness about the data and choices for inclusion criteria in published meta-analyses has been raised as one of the problems in resolving debates about meta-analyses (e.g., Bushman, Rothstein, & Anderson, 2010). Making all meta-analytic data publicly available facilitates continuously cumulating meta-analyses that update effect size estimates as new data are collected (e.g., Braver, Thoemmes, & Rosenthal, 2014), as new theoretical viewpoints emerge, and as new methods to correct for publication bias in meta-analyses are developed. For example, novel statistical techniques (*p*-curve, *p*-uniform, and meta-regression) aim to provide meta-analytic effect size estimates that correct for publication bias (Simonsohn, Nelson, & Simmons, 2014; Stanley, 2017; van Assen, van Aert, & Wicherts, 2015), but these techniques can only easily be applied to reproducible meta-analyses. Unfortunately, meta-analyses can rarely be reproduced. Gøtzsche, Hróbjartsson, Marić, and Tendal, 2007 examined whether standardized mean differences in meta-analyses of medical research were correctly calculated and reproducible for 27 meta-analyses. In 10 (37%) of these meta-analyses, results could not be reproduced within a predefined difference of the effect size estimate of Hedges $g = 0.1$. Among

these ten meta-analyses, one that had yielded a significant meta-analytic effect was no longer significant, three without significant results yielded significant results when reproduced, and one meta-analysis was retracted.

The current project aimed to take the lessons we have learned from recent theoretical, methodological, and replication papers about challenges in individual studies, and applies these lessons to understanding the strengths and weaknesses of current practices when performing and publishing meta-analyses. We set-out to reproduce 20 recently published meta-analyses. Our goals were to (1) assess the similarity between published and reproduced meta-analyses and identifying specific challenges that thwart reproducibility; (2) examining the extent to which current best practices (e.g., PRISMA (Moher, Liberati, Tetzlaff, & Altman, 2009), MARS (Publication Manual of the American Psychological Association, 2009) are adhered to, and how these guidelines can be improved to insure the replicability of future meta-analyses relying on novel statistical techniques, and; (3) investigate the extent to which novel statistical techniques that correct for publication bias impact meta-analytic effect size estimates, and thus the importance of facilitating the application of such new techniques.

Methodology

We collected all (54) meta-analyses that were published in *Psychological Bulletin, Perspectives on Psychological Science, and Psychonomic Bulletin & Review* in 2013 and 2014. From these 54 meta-analyses, only 4 (7%) explicitly referred to the use of established reporting guidelines (all four meta-analyses used the PRISMA guidelines, Liberati et al., 2009). Only 67% or the meta-analyses (36) included a table where each individual effect size, and the study it was calculated from, was listed. The absence of such a table is, as we will discuss below, an important reason meta-analyses can not be reproduced.

Twenty out of the 54 meta-analyses were randomly selected to be reproduced. Teams attempted to reproduce these meta-analyses as completely as possible based on the specifications in the original articles. As part of this project, we developed a coding scheme that should facilitate the reproducibility of meta-analyses, while allowing all currently available meta-analytic techniques to be applied to the data. This includes p-curve analysis (Simonsohn et al., 2014), p-uniform (Aert, Wicherts, & Assen, 2016), statcheck, (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2016), GRIM (Brown & Heathers, 2016), and PET-PEESE (Stanley, 2017). The coding form, and all other materials, is available from https://osf.io/q23ye/.

The difficulties we observed in reproducing meta-analyses were so severe, that some of the goals of this project could not be achieved. Therefore, in this report we will focus on our main research question: *How difficult is it to reproduce published meta-analyses in psychology?* Our second research question about the impact of subjective choices in meta-analyses is discussed qualitatively, but due to the challenge to reproduce

meta-analyses to begin with, we can not address this question quantitatively. Finally, our last goal to apply novel meta-analytical tools to existing meta-analyses. Although we plan to examine this question for the few meta-analyses that could be quite accurately reproduced, addressing this question was not possible for most meta-analyses in this project. The goal of applying these new techniques was to compare the bias-corrected effect size estimate to the effect size estimate reported in the meta-analysis, but for most meta-analyses it was not possible to reproduce enough effects with certainty that a direct comparison was possible. Nevertheless, the coding scheme for meta-analyses that was developed during this project is an important contribution in itself, and should enable the application of these novel techniques in the future.

We want to stress that we do not believe the authors who have published meta-analyses are to blame for the difficulty in reproducing their published meta-analyses. Some challenges in reproducing meta-analyses, such as the limited availability of unpublished data, and the lack of public availability of data underlying published datasets, make it difficult to by default reproduce each effect size included in a study. These are systemic problems in closed science, which can only be solved by embracing open data. Current standards in psychology don't require authors to publish reproducible meta-analyses, nor do they require authors to follow guidelines such as PRISMA or MARS. There is considerable variability in how meta-analyses are reported, and when reporting guidelines are not followed, it is simply not required to publish reproducible meta-analyses. Because there is, as far as we know, no research on how reproducible meta-analyses in psychology currently are, we expect neither authors, reviewers, or editors are aware of the difficulties in reproducing meta-analyses. We hope this project will instigate a discussion about whether it is desirable that meta-analyses are reproducible in psychology, and how this goal can be achieved.

Coding Meta-Analyses

How should researchers code the scientific literature in their meta-analytic research to facilitate reproducibility and transparency? At the start of this project, we developed a coding scheme that should be specific enough to guarantee all relevant information is coded from the primary literature, but broad enough to be applicable to all meta-analyses in psychology. We believe the final solution contains some worthwhile improvements to current recommendations, and we will discuss these by comparing them to the recommendation in the Handbook of Research Synthesis (Cooper, Hedges, & Valentine, 2009), especially as described in chapter 9 by David Wilson on systematic coding.

Identify Report. An individual article is a logical unit of primary information. Each individual article should be identified by name. Where Wilson suggests to code "basic information about the manuscript, such as

the authors' names, type of manuscript (journal article, dissertation, technical report, unpublished manuscript, and so on), year of publication, and year or years of data collection" we chose to code each article based on its Digital Object Identifier, or DOI. The DOI can be used to automatically retrieve the basic information suggested by Wilson, and more, and therefore is the preferred identifier when identifying individual articles while coding meta-analyses. Although a DOI is easily recognized digitally, it is not easy to recognize the study it identifies for the eta-analyst, and this adding a column that simply includes the article name (e.g., "Lakens, 2014") is helpful while coding meta-analyses.

Identify Cohort. One aspect currently absent from most (or perhaps even all) recommendations on how to code meta-analyses is to code which participant sample is used in the data analysis. This is important when the same sample is measured at different times (e.g., in a longitudinal study), or when data of the same sample is reported across different publications. For example, in the meta-analysis on gender differences in emotion expression Chaplin and Aldao (2013) include four effect sizes from Buss (2011) and one effect size from Brooker and Buss (2010). All four effect sizes from Buss (2011) are from the same participants, and coding these as four different effect size estimates inflates the sample size three times, since the data is from the same cohort. Moreover, even the data from Brooker and Buss (2010) is from this same cohort of young children, and it is therefore erroneous to treat this effect size estimate as independent from the other four. All these effects can be coded separately, but including a cohort DOI (i.e., the DOI of the study in which this cohort was used for the first time) will force meta-analysts to explicitly consider how the effect size estimates from the same group of people will be combined in a meta-analysis.

Identify Study. In multiple study papers, it is useful to identify the study an effect size was coded from.

Study Design. In perhaps the most proto-typical research scenario, participants are randomly assigned to either an experimental condition, or a control condition. But in other areas in psychology, within designs are more common. Whether a research hypothesis was tested in a within or between design is important when calculating effect sizes, and thus, this information should be coded.

Result. Effect sizes are calculated from results. Wilson recommends coding both the effect size, as the raw data upon which this effect size is calculated. Furthermore, Wilson notes: "it is useful to record the page number of the manuscript where the effect size data can be found". However, of all meta-analyses we reproduced in our project, there was no meta-analysis that provided either the page, nor the data upon which the effect size was calculated. In the 67% of meta-analyses where each individual effect size was listed, only the effect sizes and sample sizes were provided. Very often, we were not able to determine which data was used to

calculate the effect size. We therefore recommend to explicitly code the result by copy-pasting the text upon which the effect size is based (e.g., “There was a main effect of block, (M = 52, SD = 6, vs. M = 65, SD = 8), $F(1, 65) = 4.13$, $p < .05$”). This will greatly facilitate the identification of the data used to code the effect size, and thus the reproducibility of meta-analyses.

(Effect Size) Data. Most meta-analyses are performed based on effect sizes, and thus, general recommendations focus on the information that is needed to calculate effect sizes. Indeed, effect size data is an important source of information, but not the only source of information researchers might want to code. Additional meta-analytical techniques exist, such as p-curve analysis and p-uniform, in addition to techniques to check for errors in the reported statistics, such as statcheck and GRIM. In our coding scheme, we have made all information needed to calculate effect sizes explicit, and include information needed for additional analyses. We believe researchers should aim to code all this meta-data, and share it with their meta-analysis.

Among the information that is needed is the sample size used in the analysis (and when independent groups are part of the design, the sample size in the two groups that are compared), for t-tests and F-tests the degrees of freedom, a specification of the test statistic that is reported (which has a wide range of possibilities, such as a correlation, t-statistic, Z-statistic, chi-squared statistic, unstandardized beta, etc.) and the test value. For statcheck, the direction of the reported comparisons (e.g., <. =, >) and the reported p-value is required, and for GRIM (and in general, comparisons between means) the means and standard deviations should be coded. We recommend researchers to explicitly code the correlation between dependent means for within designs, and numbers of failures and successes for each group for dichotomous outcomes.

Effect Size Formula. For each reported effect size researchers should specify, in detail, which effect size formula is used. General references to books that present effect size formula’s is not specific enough, but references to specific page numbers are specific enough, as long as these pages only contain a single relevant equation. Without effect size formula, it is often not easy to reproduce the effect size, even when the effect size data can be identified.

Comments. Finally, a comment field is often required to specify assumptions that were made, or justify subjective choices. In many research lines, several justifiable choices for effect sizes can be made. There is no other way to communicate how these subjective choices were made. It is often impossible to develop a set of rules that succinctly summarize all the choices that need to be made when coding effect sizes, which assumptions have to be made, etc.

Moderators. Because an important goal when performing meta-analyses is to explain observed heterogeneity by moderators, researchers should code possible moderators. There is a wide range of possible moderators, and plausible moderators depend on the dominant theories in a given research area. In our project, we did not reproduce the coding of moderators, given that evaluating these choices requires more extensive knowledge about the research domain than time permitted in the current project.

Preliminary Observations

Five meta-analyses proved so difficult to reproduce that is was not deemed worthwhile to continue with the attempt. For these five meta-analyses, two did not contain a table with all effect sizes, making it impossible to know which effects were coded, and to check whether effects were coded correctly. Two other meta-analyses could not be reproduced from articles in the published literature because the relevant information was missing, and thus needed access to raw data, and in another meta-analysis not enough details were provided on how effect sizes were selected and calculated. Coding for 9 meta-analyses was completed, and coding for 6 meta-analyses is completed to some extent, but still ongoing. All reproduced meta-analyses can be found on the Open Science Framework: https://osf.io/q23ye/files/. Overall, there was unanimous agreement across all seven teams involved in extracting data from the literature that reproducing published meta-analyses was much more difficult than we expected.

Reproducibility of effect sizes was optimal when the meta-analysis reported a table specifying all effect sizes included in the meta-analysis, and when the studies included in the meta-analysis had a relatively homogenous reporting style for the primary outcome. For example, in a meta-analysis by Liu and colleagues (Liu, Huang, & Wang, 2014) the authors examined whether job training interventions had a positive effect on the percentage of people who were employed several months later. In this literature, almost all results included in the meta-analysis consisted of the number of people in the experimental and control conditions who were or were not employed. This result was almost always clearly reported in the study, often in the abstract of research reports. In this meta-analysis 47 effect sizes were reported. Ten research reports included in the meta-analysis could not be found in the literature (e.g., because the data was from a dissertation or thesis). A total of 37 effect sizes could be calculated for the remaining articles. Following (Gøtzsche et al., 2007) we considered effect size calculations within a 0.1 range successful reproductions. Six out of 37 effect sizes in this meta-analysis were not successfully reproduced (16%). In this meta-analysis job search success was the main dependent variable, and data was typically available for multiple measurement times (e.g., three months after the intervention, 6 months after the intervention, and a year after the intervention). In two cases, the difference in the calculated effect size

might have been due to difficulty selecting the correct measurement time, and one difference was caused by a different choice for the best control group. In another instance, we chose to use the effect size based on all participants (including drop-outs), which would be consistent with other effect size calculations in the same meta-analysis, but for this specific study, the authors of the meta-analysis used an effect size based on the data without drop-outs from the study (so only a subset of all participants). In one case an error in the calculation of a percentage in the original article was copied into the meta-analysis, where we corrected the error in the original research report, which changed the effect size estimate in the meta-analysis. In another case, an error in the original article (with conflicting information in the text and a table) led to a different effect size estimate because we made a different judgment about which numbers were correct than the authors of the meta-analysis. Thus, for this meta-analysis, 4 out of six differences were due to subjectively justifiable choices, and two were due to a different treatment of errors in the published articles.

Another example where reproducibility was very high was in the meta-analysis by Klahr and Burt (2014) who examined genetic influences on parenting behavior. Because intraclass Pearson product–moment correlations are the default measure reporting in practically all the original studies, it was relatively straightforward to reproduce this meta-analysis. However, even there, one transcription error was found.

Regrettably, not all meta-analyses were of this high quality. For example, when reproducing the meta-analysis by Chaplin and Aldao (2013) it became clear that the authors calculated multiple effect sizes from the same participants, and combined these in meta-analytic effect size estimates even though these effect sizes are dependent. As indicated above, introducing a Cohort ID, where researchers need to report the sample or cohort the effect size is based on, should make it clear that effect sizes are dependent, even in more subtle cases where data from the same sample is reported in different articles published in different years. Since it is not very informative to reproduce a faulty meta-analysis, coding for this meta-analysis was not completed, but even in the 30 effect sizes that were reproduced before the errors in the original meta-analysis became clear, two error ware identified (one in the sample size, one in the calculation of the effect size), four reproduced effects differed considerably from those calculated in the meta-analysis, 16 effects could not be reproduced because the raw data was needed, and only eight effect sizes were successfully reproduced.

Many effect sizes in meta-analyses proved impossible to reproduce due to a combination of a lack of available raw data, difficulty locating the relevant effect size from the article, unclear rules about how multiple effect sizes were combined or averaged, and lack of specification of the effect size formula used to convert effect sizes. The difficulty in locating the correct effect size can be due to the fact that our team lacked expertise

in a specific area, although many times, even when calculating effect sizes for all reported results, none matched the effect size reported in the meta-analysis. It would be easier to identify relevant results when they are explicitly described (e.g., copy-pasted from the text, or referred to by table number and line in the table) in the coding form. Effect size equations used to convert effect sizes should be provided for each effect, and when more information is needed about how relevant information was retrieved (e.g., when data was estimated from a figure), this should be described in the comments section of the coding form. It is difficult to know why reported and reproduced effect sizes differed, but it highlights that it is often not very clear how effect sizes in a meta-analysis were calculated. This is undesirable, because it reduces the transparency of meta-analyses. Furthermore, coding errors happen, and when effect sizes can not be reproduced, it is almost impossible to check and correct meta-analysis.

Although the meta-data coding form was rather complete, evaluation after the data coding yielded possible improvements. Wilson, in chapter 9 of the handbook of research synthesis (Cooper et al., 2009), recommends to explicitly code whether the observed effect is in the direction of the research hypothesis (1 = favors treatment, 2 = favors control, 3 = neither). In the current form, coders were expected to code effect sizes in a specific direction (often indicated in the meta-analysis). However, this makes this aspect of the coding less transparent than when the direction of the effect is coded explicitly. Although in psychology studies might not be coded as easily as 'treatment vs control', it seems preferable to explicitly code the direction of the effect. As an example of why this is important, one meta-analyst treated the direction of an effect inconsistently for two different predictors in her own paper (a negative effect for one predictor was coded as a positive effect size, but not for another predictor, suggesting inconsistency in reverse scoring). By explicitly coding the direction of the effect in the coding form, these choices become more visible.

A frequent source of disagreement between the meta-analysis and the recoded effect sizes concerned the sample size in the study, or the sample sizes in each group. Because sample sizes are used to calculate the variance for effect sizes, this is important information to agree on, but due to the fact that sample sizes for specific analyses are often not clearly provided in original articles, this information turned out to be surprisingly difficult to code. Original research article could take greater care in reporting this information for analyses, especially when not all participants who take part in a study end up in the final data analysis.

Improving the Reproducibility of Meta-Analyses

PRISMA and MARS guidelines provide many excellent recommendations, and researchers who perform a meta-analysis should carefully consider following these guidelines. As noted before, only 67% of the

54 meta-analyses that were published in Psychological Bulletin, Perspectives on Psychological Science, and Psychonomic Bulletin & Review in 2013 and 2014 included a table that specified each coded effect size. Some meta-analyses provided a forest plot where each effect size and confidence interval is visualized. This might be due to PRISMA guidelines that stress that effect sizes and confidence intervals should be reported, but seem to give authors the option to either report this information in a table or in a plot ("This information may be incorporated in a table showing study characteristics or may be shown in a forest plot"). Although PRISMA guidelines note it is 'preferable' to report numeric data, we recommend authors follow the MARS guidelines, which clearly specify to report a "Table giving descriptive information for each included study, including effect size and sample size". The numerical effect sizes, sample sizes, and confidence interval should be reported in 100% of meta-analyses to facilitate reproducibility.

We believe authors should share all information they coded or calculated when performing the meta-analysis (Lakens, Hilgard, & Staaks, 2016). The meta-data coding spreadsheet we developed for the current project might be a good starting point for a default solution for meta-analyses in psychology. There is no doubt room for improvement, but the fact that the current coding scheme sufficed for all 20 meta-analyses we aimed to reproduce in the current project indicated it should cover the basic information (but not the moderators) needed to code meta-analyses across a wide range of research areas in psychology. There are substantial benefits to using a default coding form across psychology, whenever possible. The meta-data can be easily combined in larger data-bases (for an example, see Metabus, Bosco, Steel, Oswald, Uggerslev, & Field, 2015) thus generating a huge collection of meta-data that can be used for a wide range of applications. Currently, all individual meta-analyses represent a huge amount of coding effort, but this meta-data is not made available in any useful way.

When extracting information from the literature, meta-analysts often ask researchers for raw data underlying their articles, to code effects not reported in the published manuscript. Despite APA Journal Article Reporting Standards (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) instructing researchers to report "Direction, magnitude, degrees of freedom, and exact *p* level, even if no significant effect is reported" many academic psychologists have the bad habit of reporting non-significant results in such a way that no effect size can be calculated (e.g., $t(122) < 1$, n.s, or $F(2,420) = 0.98$, $p > 0.05$).

For example, Chaplin and Aldao (2013) retrieved 52 datasets from authors who published research on emotion expression in children to be able to calculate gender differences in emotion expression. As the authors

note: "*For papers lacking sufficient information that were conducted in the past 12 years (from 1999 to 2010; N = 209), we requested data from the authors. Authors provided data for 52 of these studies. As in other meta-analysis studies (e.g., Else-Quest et al., 2006), data were not requested for articles published prior to the past 12 years, as these data are known to be difficult to retrieve.*" These authors make explicit that their meta-analysis will not be reproducible the moment it appears in print. If we take the idea that we can only expect to retrieve data from articles published up to 12 years ago seriously, the moment their meta-analyses is published (2013) datasets from 1999 to 2002 are no longer available for anyone who wants to reproduce the meta-analysis. This is an extremely undesirable situation. For our current project, meta-analyses where a large percentage of effect sizes could not be calculated from the literature were considered not reproducible. We should state explicitly that we did not attempt to contact the researchers who published the original research in this project, but determined an effect was not reproducible when the effect size could not be coded based on the study that was included in the meta-analysis. One reason for this choice was that it was not feasible to collect the raw data for all the studies where the effect size could not be calculated based on the original article for all 20 meta-analyses. Furthermore, it was often not clear for which effect sizes meta-analysts relied on data they received from authors. It would be extremely useful if meta-analysts did not just indicate that they received raw data from authors, or how many effect sizes were based on raw data retrieved from authors, but indicated each effect size that was based on raw data that was retrieved from original researchers.

From the earlier example, it is clear that the only true solution to cumulative science and reproducible meta-analyses is for researchers to share their data with their publication, whenever this is possible. It is well known researchers often do not share their data when fellow researchers ask for it. For example, in the meta-analysis by Chaplin and Aldao (2013) 52 datasets were received from 209 publications, which means only 25% of the requested datasets were made available. This low rate of data sharing seems to be a general principle in psychology (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; Wicherts, Borsboom, Kats, & Molenaar, 2006), and there are indications the availability of data decreases over time (Gibney & van Noorden, 2013). Although data-sharing with raw publications was not typical in the past, data-sharing is becoming more common in psychology (Kidwell et al., 2016), and some have argued data-sharing should be the default, with exceptions being justified (Morey et al., 2016). Given these changing norms, as well as the technological feasibility to easily share data, collecting raw data from original researchers is an opportune moment to ask these researchers if you have their permission to make their data publicly available with the published meta-analysis.

Datasets needed when research reports do not report sufficient information can be difficult to retrieve, but so can research reports themselves, especially when research is reported in theses, conference posters, or unpublished manuscripts. In total, the 20 meta-analysis randomly chosen for coding extracted effect sizes from 2080 articles. We searched the literature to retrieve as many articles as possible, but we were unable to locate 225 studies (11%). These missing articles consisted partly of theses (some of which could be downloaded through ProQuest, which we did not have access to at Eindhoven University of Technology), unpublished datasets, and conference proceedings. Whenever copy-right allows it, meta-analysists could share research reports used in the meta-analysis, for example in a folder on the Open Science Framework.

Conclusions

This project had three aims. First, we wanted to examine how difficult is it to reproduce published meta-analyses in psychology. Second, we wanted to know what the impact of subjective choices on the meta-analytic effect size estimates was. And finally, we were interested in whether conclusions change when novel statistical techniques are applied to published datasets. In this preliminary report, we show the great difficulty we faced while attempting to reproduce 20 meta-analyses published in three psychology journals in 2013 and 2014. We believe this is a situation in dire need of improvement.

We want to highlight that none of the authors of the meta-analyses we have reproduced have deviated in any way from the current standards that are expected in top journals in psychology when publishing meta-analyses. Reproducibility is simply no requirement for publication. Making meta-analyses reproducible requires careful thought and planning, and is a skill that requires expertise that in all likelihood needs to be developed over the next few years. Our work here might be considered a first step on the way to more transparent and reproducible meta-analyses in psychology.

Any researcher who performs a first meta-analysis will probably recognizing sub-optimal reporting practices in their own empirical articles in the past. Doing a meta-analysis is a great way to learn which information one needs to report to make sure the article can be included in a future meta-analysis. Similarly, reproducing a meta-analysis is a great way to learn which information one needs to report in a meta-analysis to make sure a meta-analysis can be reproduced. We can highly recommend researchers interested in performing a meta-analysis to start out by (in part) reproducing a published meta-analysis.

The lack of adherence to PRISMA or MARS reporting guidelines when meta-analyses in psychology are published is one of the most important improvements that is straightforward to implement. Adhering to reporting guidelines is a minimum requirement to get published in many other research domains, such as health

research, where a wide variety of reporting guidelines are commonly used (see the equator website for a comprehensive collection: https://www.equator-network.org/reporting-guidelines/). Ideally, every meta-analysis that is published in psychology should follow either the MARS or PRISMA guidelines. Furthermore, a single coding scheme for all meta-analyses that is shared with each published meta-analysis should make it easy to combine information across meta-analyses, over time leading to a huge database of hand-coded meta-data that can be used for scientific purposes.

As long as the reproducibility of meta-analyses is not improved, it is difficult to compare the impact of subjective choices on the meta-analytic effect size, or to apply novel meta-analytic techniques to published meta-analyses. It was relatively common to see a difference between an effect size or sample size reported in a meta-analysis, and the effect size of sample size that was recoded. This difference might be caused by the fact that the person reproducing a meta-analysis had less expertise on a specific topic than the researchers who performed the meta-analysis (although relatively little expertise is needed to code a sample size, and some clear mistakes were observed). Different coding choices might be defensible, and have relatively little impact on the meta-analysis overall, but they should be transparent, which is not the case in published meta-analyses.

Current reporting guidelines for meta-analyses were not developed to guarantee reproducibility. Our project clearly demonstrates the need to develop dedicated guidelines to improve the reproducibility of meta-analyses, such that researchers can to easily check and update meta-analyses in the future. We have discussed a number of recommendations in this project, such as clearly identifying and coding the data each effect size is based on, providing detailed explanations how multiple effect sizes from the same study are combined, and explaining which effect size calculations are used. Meta-analyses are typically regarded as important sources of information, and many published meta-analyses are cited often. Because meta-analyses play an important role in cumulative science, they should be performed with great transparency, and be reproducible. With this project we hope to have provided some important preliminary observations that stress the need for improvement, and provided some practical suggestions to make meta-analyses more reproducible in the future.

References

Aert, R. C. M. van, Wicherts, J. M., & Assen, M. A. L. M. van. (2016). Conducting Meta-Analyses Based on p Values Reservations and Recommendations for Applying p-Uniform and p-Curve. *Perspectives on Psychological Science*, *11*(5), 713–729. https://doi.org/10.1177/1745691616650874

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Bosco, F. A., Steel, P., Oswald, F. L., Uggerslev, K., & Field, J. G. (2015). Cloud-based Meta-analysis to Bridge Science and Practice: Welcome to metaBUS. *Personnel Assessment and Decisions*, *1*(1). Retrieved from https://scholarship.rice.edu/handle/1911/87848

Braver, S., Thoemmes, F., & Rosenthal, R. (2014). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Brooker, R. J., & Buss, K. A. (2010). Dynamic measures of RSA predict distress and regulation in toddlers. *Developmental Psychobiology*, *52*(4), 372–382. https://doi.org/10.1002/dev.20432

Brown, N. J. L., & Heathers, J. A. J. (2016). *The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology* (No. e2064v1). PeerJ Preprints. Retrieved from https://peerj.com/preprints/2064

Bushman, B. J., Rothstein, H. R., & Anderson, C. A. (2010). Much ado about something: Violent video game effects and a school of red herring: Reply to Ferguson and Kilburn (2010). Retrieved from http://psycnet.apa.org/psycinfo/2010-03383-004

Buss, K. A. (2011). Which fearful toddlers should we worry about? Context, fear regulation, and anxiety risk. *Developmental Psychology*, *47*(3), 804–819. https://doi.org/10.1037/a0023227

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin*, *139*(4), 735–765. https://doi.org/10.1037/a0030737

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed). New York: Russell Sage Foundation.

Ferguson, C. J. (2014). Comment: Why meta-analyses rarely resolve ideological debates. *Emotion Review*, *6*(3), 251–252.

Gibney, E., & van Noorden, R. (2013, December 19). Scientists losing data at a rapid rate. Retrieved March 30, 2017, from http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences. *JAMA*, *298*(4), 430–437. https://doi.org/10.1001/jama.298.4.430

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

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klahr, A. M., & Burt, S. A. (2014). Elucidating the etiology of individual differences in parenting: A meta-analysis of behavioral genetic research. *Psychological Bulletin*, *140*(2), 544–586. https://doi.org/10.1037/a0034205

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychology*, *4*, 24. https://doi.org/10.1186/s40359-016-0126-3

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, *339*(jul21 1), b2700–b2700. https://doi.org/10.1136/bmj.b2700

Liu, S., Huang, J. L., & Wang, M. (2014). Effectiveness of job search interventions: A meta-analytic review. *Psychological Bulletin*, *140*(4), 1009–1041. https://doi.org/10.1037/a0035923

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … others. (2016). The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … others. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. https://doi.org/10.1126/science.aac4716

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555.

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.

Stanley, T. D. (2017). Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science*, 1948550617693062. https://doi.org/10.1177/1948550617693062

van Assen, M. A. L. M., van Aert, R. C. M., & 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

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra: Psychology*, *1*(1). https://doi.org/10.1525/collabra.13

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726.