Reproducibility of published meta-analyses on clinical psychological interventions

Rubén López-Nicolás[1], Daniel Lakens[2], Jose A. López-López[1], Maria Rubio-Aparicio[3], Alejandro Sandoval-Lentisco[1], Carmen López-Ibáñez[1], Desirée Blázquez-Rincón[1], & Julio Sánchez-Meca[1]

[1] University of Murcia, Spain

[2] Eindhoven University of Technology, The Netherlands

[3] University of Alicante, Spain

Author Note

Correspondence concerning this article should be addressed to Rubén López-Nicolás, Facultad de Psicología, Campus de Espinardo, Universidad de Murcia, edificio n° 31, 30100, Murcia, Spain. E-mail: rlopez@um.es

Abstract

Meta-analysis is one of the most useful research approaches, the relevance of which relies on its credibility. Reproducibility of scientific results could be considered as the minimal threshold of this credibility. We assessed the reproducibility of a sample of meta-analyses published between 2000-2020. From a random sample of 100 papers reporting results of meta-analyses of interventions in clinical psychology, 217 meta-analyses were selected. We first tried to retrieve the original data by recovering a data file, recoding the data from document files, or requesting it from original authors. Second, through a multi-stage workflow, we tried to reproduce the main results of each meta-analysis. The original data were retrieved for 67% (146/217) meta-analyses. While this rate showed an improvement over the years, in only 5% of these cases was it possible to retrieve a data file ready for reuse. Of these 146, 52 showed a discrepancy larger than 5% in the main results in the first stage. For 10 meta-analyses this discrepancy was solved after fixing a coding error of our data retrieval process and for 15 of them it was considered approximately reproduced in a qualitative assessment. In the remaining meta-analyses (18%, 27/146), different issues were identified in an in-depth review, such as reporting inconsistencies, lack of data, or transcription errors. Nevertheless, the numerical discrepancies were mostly minor, with little or no impact on the conclusions. Overall, one of the biggest threats to the reproducibility of meta-analysis is related to data availability and current data sharing practices in meta-analysis.

Keywords: meta-analysis, reproducibility, data sharing, data reusability, research synthesis

Reproducibility of published meta-analyses on clinical psychological interventions

Meta-analysis is widely considered as an important approach to evaluate a body of work. Given the ongoing growth in the number of scientific publications (Bornmann et al., 2021), evidence synthesis approaches—such as meta-analysis—are becoming increasingly relevant for a cumulative science. This relevance rests on the credibility of meta-analytic results, which can be threatened by a lack of rigorous methodology or poor-quality reporting (Gurevitch et al., 2018). Given the importance of meta-analyses for evidence-based practice, these threats to their credibility need to be closely monitored.

In recent years different concerns on the credibility of empirical claims have emerged. Several projects have systematically attempted to assess the replicability and reproducibility of published scientific results (e.g., Artner et al. (2020); Errington et al. (2021); Open Science Collaboration (2015)). Those initiatives showed many failures to replicate or reproduce the published results. In this context, the empirical assessment of the credibility of published results has become a major task for the scientific community.

There are different approaches to the empirical assessment of scientific credibility. Reproducibility refers to the attempt to obtain the same results as in the original publication, using the same data and the same procedure. Robustness refers to the assessment of the sensitivity of the originally published results and conclusions to variations in the original procedure using the same data. Replicability is a core principle of the scientific method and refers to the fact that the same scientific evidence should be observed when independent researchers try to answer the same research question from the same approach at different moments using different data. In other words, obtaining the same results, using different data and answering the same question (National Academies of Sciences, Engineering, and Medicine, 2019; Nosek et al., 2022). In this project, we focus on the reproducibility of meta-analyses.

The reproducibility of published scientific results could be considered as the minimal

threshold of scientific credibility (Hardwicke et al., 2021). Different approaches can be adopted for the empirical assessment of reproducibility. For example Nosek et al. (2022) make the distinction between process reproducibility and outcome reproducibility. Following this framework, a process reproducibility assessment could be carried out by reviewing the availability of the materials, data, or precise details of the analytical strategy in the report that are required to proceed with the reproduction attempt. An outcome reproducibility assessment can be carried out when the required elements are retrievable by actually reproducing the analyses. It is worth noting that the difficulty of performing an outcome reproducibility assessment depends on which analytical information is available. The availability of the original analysis code (i.e., the original computational instructions in a programming language) facilitates reproducibility analysis by enabling simply re-running the code on the data. Regrettably, the analysis code is currently seldom available (Hardwicke et al., 2020, 2022; López-Nicolás et al., 2022). When only a verbal summary of the performed analyses is available in the research report (which is the most common scenario in practice), the original analysis needs to be reconstructed. The challenges and implications of failed reproductions in both cases may be of a different nature.

Several reproducibility analyses of meta-analyses have been performed in recent years. For example, some process reproducibility assessments have shown an important lack of data availability in machine-readable formats, and an almost complete absence of analysis script code availability (López-Nicolás et al., 2022; Polanin et al., 2020). Furthermore, some outcome reproducibility assessments have shown a considerable number of failures when trying to reproduce the primary effect sizes of some published meta-analyses by recollecting primary data from primary studies (Gøtzsche et al., 2007; Maassen et al., 2020; Tendal et al., 2009), possibly due to lack of details on how primary effect sizes were selected and computed. In these outcome reproducibility studies, the main task entails reconstructing the original data by retrieving them from the source, namely the included primary studies. Thus, their assessment focus is on this stage of the analysis pipeline of a meta-analysis, which usually

involves decisions on how to select the primary outcomes and how to deal with possible dependency, and the computation of (standardized) effect sizes. Figure 1 displays a summary of the basic meta-analysis pipeline through a flowchart, outlining the different stages and listing previous work that has explored different facets of reproducibility of these, as well as a summary of the required elements to be able to reproduce each stage. In this project we focus on the last stage, related to the statistical analysis and quantitative results of the synthesis.

Reproducibility analysis of reported quantitative results typically uses the original data available from the original authors (e.g., Artner et al., 2020; Hardwicke et al., 2018, 2021). This puts the focus of the assessment at factors such as the reusability of the available data, challenges for the reconstruction of the original analysis scheme, reporting errors, etc. Although data availability seems to have improved in the last years (Hardwicke et al., 2018; Tedersoo et al., 2021; Wallach et al., 2018), systematic reviews and meta-analyses appear to be a special case. Typically, the data collected for a meta-analysis is study-level summary data extracted from published primary studies which is commonly reported in the paper through tables or forest plots. This may lead to the idea that common data sharing practices do not apply to meta-analysis. For example, Page et al. (2022) analysed the content of data availability statements from a set of meta-analyses published in 2020. Only 31% included a data availability statement and only 13% of these included a link to access the data openly, with 23% stating that all relevant data are available in the paper itself, 10% stating that data sharing is not applicable as no datasets were generated, 8% stating that data sharing is not applicable as the data is drawn from already published literature, and 42% stating that data were available upon request. It is surprising that, even just considering meta-analyses that included a data availability statement, the authors of these meta-analysis assume that such practices do not apply to meta-analyses, or that the data in the article itself is sufficient.

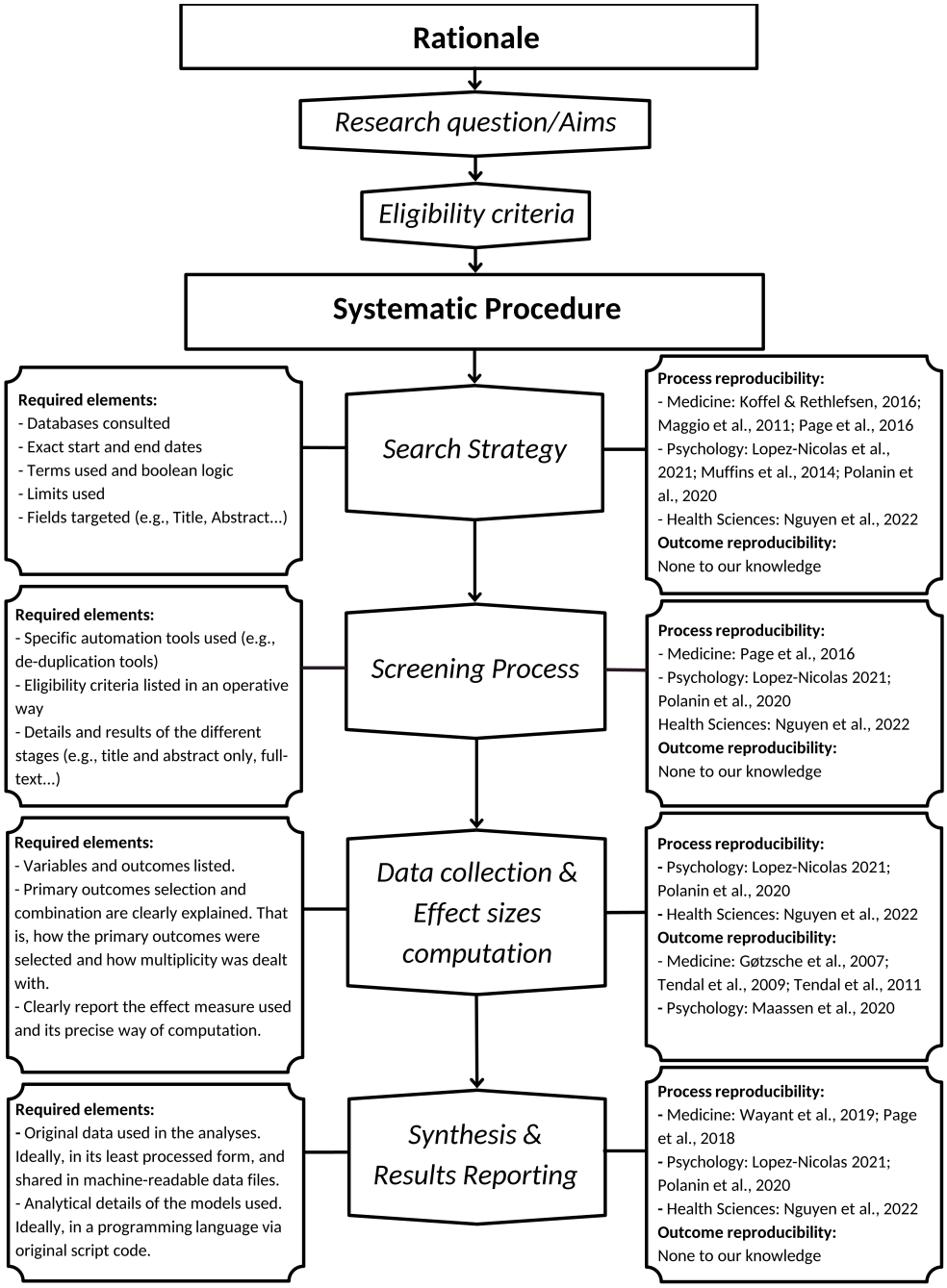

Figure 1

Flowchart displaying the basic pipeline of a meta-analysis. Each of the stages may be subject to reproducibility evaluation. On the left, known studies that have evaluated some facet of the reproducibility of each stage are listed. On the right, the various elements that must be available to reproduce each stage are enumerated.

Purpose

Previous research has revealed that there is room for improvement at different stages of the meta-analytic process pipeline. In this study our purpose is twofold. First, we broadened previous process reproducibility assessments by considering data availability on request and by contacting original authors to request required information to reproduce the meta-analysis. Second, we verified the outcome reproducibility of the meta-analyses that were process-reproducible using the available data. Where previous work focused on the reproducibility of primary effect sizes by recoding data from primary studies, we explored meta-analysis outcome reproducibility using the primary data already coded by the original authors. Therefore, we attempted to retrieve the data shared by the authors of the meta-analysis.

Disclosures

Preregistration

The pre-data analysis protocol (https://doi.org/10.17605/OSF.IO/79J2T) was pre-registered on 19 October 2021. Any deviation from this protocol is explicitly acknowledged.

Data, materials, and online resources

Data and analysis script code are openly available at: https://osf.io/6cmzh/

Reporting

Below we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Method

Identification and selection of articles and meta-analyses

In previous research we identified a pool of 664 meta-analytic reports on clinical psychological interventions published between 2000 and 2020 through a systematic electronic search (López-Nicolás et al., 2022). Of this pool, 100 were randomly selected using a random number generator between 1 and the total number of meta-analyses identified. The full search strategies and a summary of the screening process are available at: https://osf.io/z5vrn/, and the workflow of the random selection process is available at: https://osf.io/cp293/. This sample size was based on our judgement of an acceptable trade-off between informativeness and feasibility. From these 100 articles, each independent pairwise meta-analytic model of aggregate data fitted on at least 10 primary studies was selected. In case no meta-analysis reported in a paper had at least 10 studies, the meta-analysis with the highest number of primary studies was selected, which was the case for 29 of the articles included in this report. This criterion was established to focus on the main meta-analyses of each paper, based on the assumption that the search strategies would be designed to maximise the number of primary studies included that were related to the main aims of the paper.

Our unit of analysis was each independent meta-analysis selected under these criteria. A total of 217 independent meta-analyses were selected.

Retrieval of primary data

In order to be able to reproduce meta-analyses of aggregate data, primary-level[1] effects sizes and their associated standard errors are required. These are generally computed from statistics retrieved from the primary studies such as means, standard deviations or sample

[1] By primary-level data we mean aggregate data from included primary studies.

sizes. We attempted to retrieve the least processed data shared by the authors of the meta-analysis. First, we sought for the statistics used to compute primary effect sizes (e.g., means, sd); second, we sought for the primary effects sizes already computed and their standard errors (or, alternatively, the sampling variances): finally, we sought for the primary effects sizes and their confidence limits, from which the standard errors were approximated as follows:

$$se_i = (\frac{UB_i - LB_i}{2z_{\alpha/2}})$$

with se_i being the standard error of the ith effect size, UB_i and LB_i the upper and lower confidence limits of confidence interval for the ith effect size, and $z_{\alpha/2}$ the $(1 - \alpha/2)\%$ percentile of the standard normal distribution (usually, $z_{\alpha/2} = 1.96$ assuming a two-sided 95% confidence interval).

On the other hand, efforts were also made to retrieve the most reusable data possible. First, we searched for machine-readable data files through links leading to third-party repositories or in supplementary material hosted by the journal. Second, we looked for available data through tables or forest plots in the meta-analytic report itself, or in supplementary material. In these cases, the primary data had to be manually re-coded to reuse it. Finally, if the primary data of a meta-analysis were not directly available after the previous steps, we attempted to obtain the data through a request to the corresponding author identified in the associated paper. We sent an initial request in June 2021 and, if there was no reply, a subsequent reminder in October 2021. This reminder was sent to a more recent alternative email address if we were able to find one. If we were unable to obtain the data through the email request, the associated meta-analysis was labelled as not process reproducible.

Reconstructing the original analytical scheme

To proceed with reproducibility attempts of the meta-analyses that were labelled as process reproducible, we first looked for the availability of the original analysis script. When

it was available, reproducibility was checked by rerunning the original script on the associated primary data. When it was not available, we tried to reconstruct the original analytical scheme using the technical details reported in the paper. Specifically, we collected information on: (a) the meta-analytic model originally assumed; (b) the weighting scheme; (c) the between-studies variance estimator; (d) the method used to compute the confidence interval; and (e) the software used to perform the meta-analysis. If any of these details about the analytical methods were not reported, but the software used was mentioned, we inferred the first four pieces of information from the default settings of the software used. If the software used was not reported, we inferred this information from the default settings of the most used software in the sample, which was *Comprehensive Meta-Analysis*. We designed this procedure to reconstruct the original analytical scheme when the original analysis script was not available instead of trying to request it from the original authors due to: (a) not necessarily all authors of included meta-analyses will actually have an analysis script to share, because many might have used point and click software, and (b) we expected analysis script availability to be very low, and requesting it would have meant sending request for virtually every paper included in our re-analysis.

Additional information about the meta-analysis was collected that is not reported in this manuscript. The full list of variables collected is available in the Protocol (https://osf.io/tq4uf/) and a Codebook describing these variables is available at: https://osf.io/ym78s/.

Data collection procedure

Data collection procedure was carried out by five of the authors. At a first pilot stage, a random sample of five articles of the total pool was independently coded by the five members and, subsequently, in a series of meetings, disagreements between the coders were resolved by consensus. Next, the initial pool of 100 included articles was split among four coders, 25 articles each. A random sample of 25 articles of the total pool was assigned to the

fifth member to carry out independent double-coding, with the goal to examine the reliability of the data collection process. Disagreements were resolved by consensus and by double-checking the original materials. Details about inter-coder agreement are reported in the Supplementary file.

Reproducibility outcomes

Each meta-analysis was labelled using the following two-level[2] reproducibility success scheme. First, each meta-analysis was labelled as: (a) process-reproducible; and (b) not process-reproducible. In our study, not process-reproducible refers to situations where we were unable to access the primary data neither through direct extraction nor upon request[3]. Second, those labelled as process-reproducible were labelled as: (a) reproducible; (b) numerical error; and (c) decision error. Similar to previous studies (Artner et al., 2020; Hardwicke et al., 2018, 2021) an index of numerical error was computed (see Protocol https://osf.io/tq4uf/). This index expressed the difference between reproduced and original values as a percentage. To avoid labelling minor numerical discrepancies related to numerical rounding as reproducibility problems, a 5% discrepancy threshold was set. Thus, a meta-analysis was labelled as 'numerical error' if it showed a discrepancy larger than 5%.[4] Finally, the label 'decision error' refers to situations where the $p_{reported}$ fell on the opposite side of the .05 boundary in relation to the $p_{reproduced}$.

We focus on reproducibility of summary effects, their confidence bounds and the result of the null hypothesis significance test. Secondarily, we also assessed reproducibility of other

[2] This hierarchy is a minor deviation from the pre-registered protocol. It is essentially the same and the results are identical. It was introduced to improve clarity.

[3] Process reproducibility, as described above, could imply a different situation if more conditions need to be met to proceed with the reproduction attempt. In our study, this is equivalent to data availability due to our design and the stage of the meta-analysis pipeline we focused on.

[4] A sensitivity analysis using other possible criteria is reported in the supplementary file.

synthesis methods such as heterogeneity statistics.

Reproducibility checks workflow

Reproducibility checks were carried out at different stages. First, through reported analytic details or script code. When the analysis script code was available, computational reproducibility was checked by rerunning the script with the available primary data. In most cases, the analysis script code was not available. Thus, in these cases we coded the analytic details as explained above to fit equivalent meta-analytic models as a function of these details using the available primary data. This analysis scheme was programmed in the R environment (R Core Team, 2022) using the *metafor* package (Viechtbauer, 2010).

Second, given that the manual recoding process is an error-prone task, some mistakes can appear. Thus, those meta-analyses labelled as numerical error and/or decision error in the previous stage were re-assessed by a different member of the team. In cases where an error was found in the originally coded results, analytic methods and/or primary data, the meta-analyses were once reproduced again and re-labelled according to the updated results. Additionally, a qualitative assessment of the meta-analyses still labelled as numerical error and/or decision error was also carried out. The same reviewers who checked for errors produced individual reports on the possible source of the discrepancy and its reproducibility was judged qualitatively by four of the other authors. This stage was a deviation from the pre-registered protocol, and made it possible to identify situations with obvious explanations, such as rounding issues, inverted signs, etc.

Additionally, for meta-analyses that remained labelled as non-reproducible, an email was sent to the corresponding author of the associated paper explaining our aims, our approach, and our results regarding his/her meta-analysis and requesting additional information that could explain the mismatch between the original reported results and the reproduced results. We tried to solve the reproducibility issues within a month after the

request and we updated the label accordingly.

Finally, the association between publication year and the possibility of retrieving the data in one of the ways conducted in this project were explored by fitting binary logistic regression models with publication year as predictor and process-reproducibility as dependent variable. We quantified the strength of the association by calculating odds ratios and 95% confidence intervals based on the profile likelihood. These exploratory analyses were not pre-registered. Details and results are reported in Supplementary file.

Results

From the 100 included papers, 217 independent meta-analyses were selected following the criteria explained above. These meta-analyses included 18.35 primary studies on average (sd = 17.25; median = 13; interquartile range = 10-19; range = 3-134), and were cited 108.39 times on average (sd = 151.00; median = 57; interquartile range = 29-128; range = 3-1036)[5]. Figure 2 displays the distribution of number of primary studies among the meta-analyses included in our sample (panel A), the publication year distribution among the papers included in our sample (panel B) as well as the citation count distribution of those papers (panel C). Original results and characteristics of these meta-analyses are available at: https://osf.io/8jzbk

Process reproducibility

Figure 3 summarizes the primary data retrieval results. Based on the availability of primary data, either retrieved directly from the paper or upon request, 146 meta-analyses (67%, see Fig. 3a) were labelled as process reproducible. Additionally, as the time span

[5] Citation counts were retrieved from CrossRef API using the *rcrossref* package (Chamberlain et al., 2023). For two cases in which CrossRef did not return data, citation counts were consulted in Google Scholar. Both queries were done on 20/03/2023.

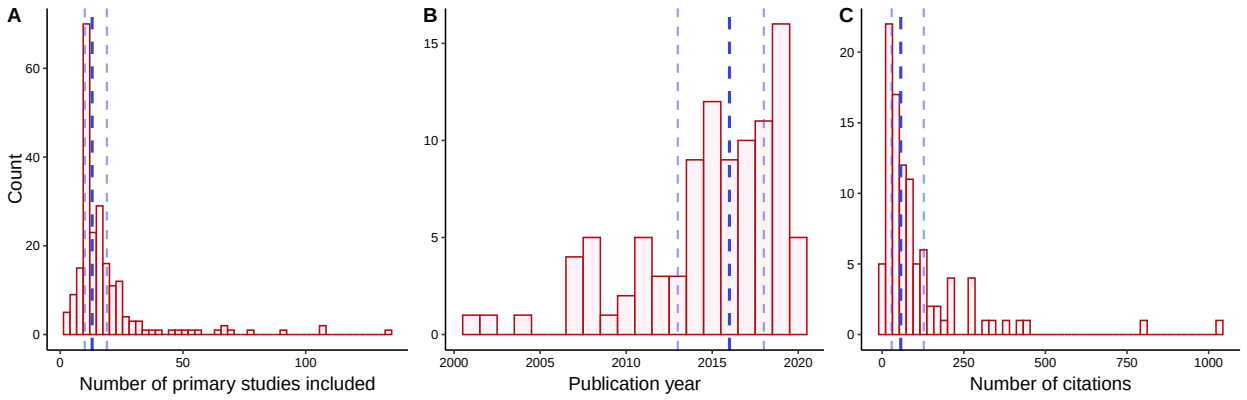

Figure 2

Distribution of (a) the number of primary studies included in each of the meta-analyses; (b) the publication year of the included papers; (c) citation count of the included papers. Vertical blue dotted lines represent the first quartile, median, and third quartile, respectively.

covered is fairly wide, the process reproducible rate was also computed for different time periods. The meta-analyses were grouped into five-year periods, except for the initial period, which was grouped into a ten-year period due to the limited number of meta-analyses available during the first five-year period, which consisted of only five meta-analyses. The process reproducibility rate was 41%, (12/29), 59%, (44/75), and 80%, (90/113) for meta-analyses published between 2000 and 2010, 2011 and 2015, and 2016 and 2020, respectively (see Fig. 3a). This trend is further explored in the Supplementary file available at: https://osf.io/fjhpw.

Of these 146 meta-analyses, in about half of the cases the primary data was retrieved from a forest plot in the paper itself and in about a third of the cases the primary data was retrieved from supplementary files (see Fig. 3b for further details). Although attempts were made to retrieve data for 75 meta-analyses from 25 different papers by emailing the corresponding authors, data was only retrieved for 4 meta-analyses, from 3 different papers (12%, 3/25, see Fig. 3c). For the remaining 71 from 22 different papers, a reply providing some reasons not to share was received in 32% (8/25, see Fig 3c), whereas no reply was received for the remainder of the meta-analyses. Table 1 summarises the different reasons

corresponding authors given when data was not provided upon request.

Table 1

Reasons given when data was not received upon request.

Reason	N
Data held by a co-author, and do not have his contact details	1
Propietary dataset	1
The author no longer has the data.	5
The author requested more information and a written agreement including possible authorship. Additional details were sent and after some email exchanges there was no further response.	1

Challenges faced retrieving primary data

In most cases, when the meta-analytic data was available, it was shared in document formats. Data were retrieved from tables or forest plots in *pdf* or *docx* format—either in the document itself or in the supplementary materials—in 92% (134/146) of the cases. This required a manual recoding of the primary data to be able to reuse them. Furthermore, when data was reported through general tables (i.e. tables listing all the primary studies included with their characteristics), the meta-analysis associated with each data entry was not always obvious, leading to the time-consuming task of matching each data entry with each independent meta-analytic result reported in the paper. There were only 4 meta-analyses (from three different papers) of the 146 meta-analyses labelled as process reproducible (3%), where the task of retrieving the data required simply downloading the data in an machine-readable data file format. On the other hand, as shown in Figure 3c, when the necessary data was not available, retrieving it upon request to the original authors led to a low response rate.

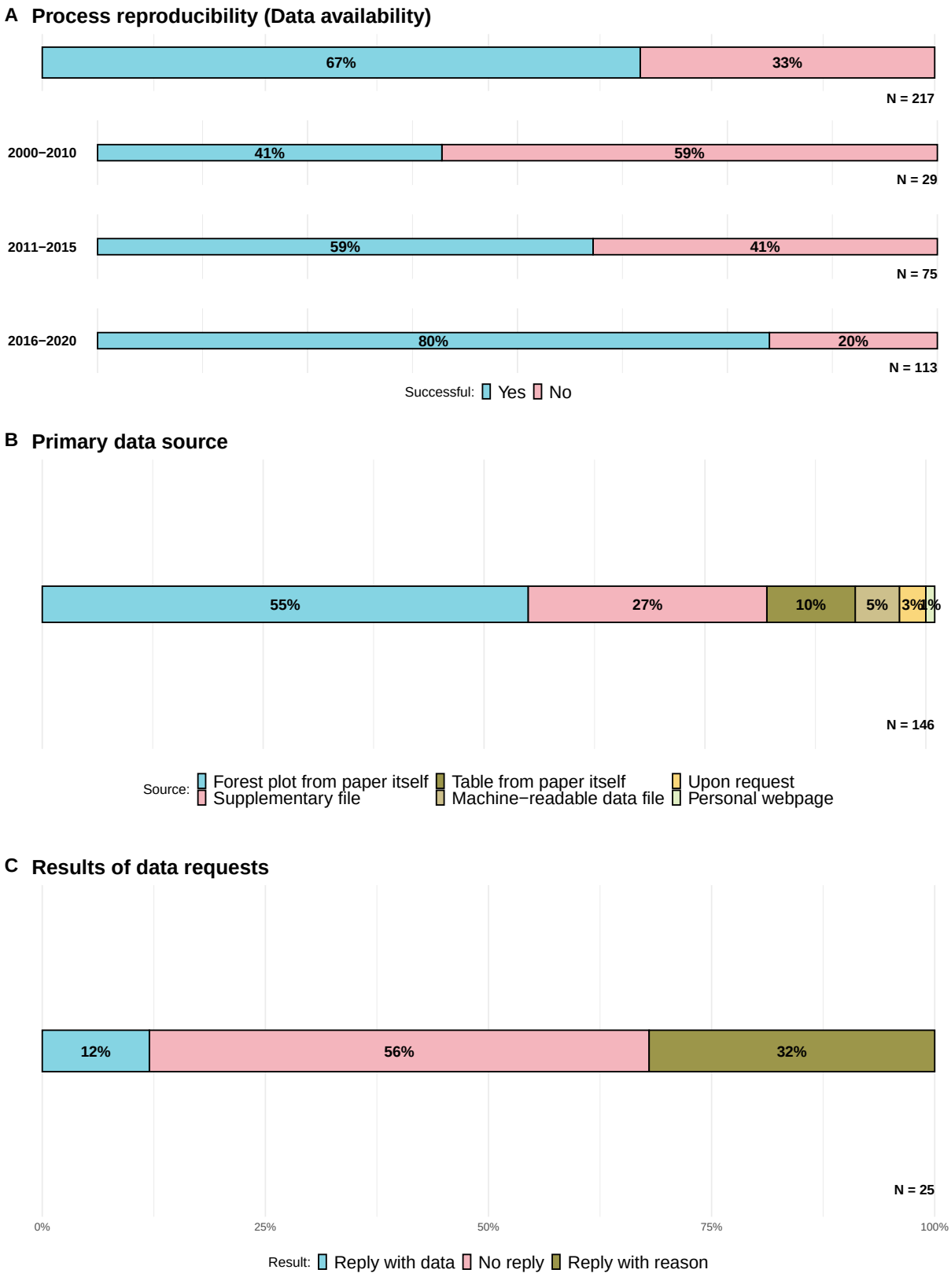

Figure 3

Percentage of (a) process-reproducible meta-analyses; (b) different types of sources of original data; (c) data request results.

Outcome reproducibility

The outcome reproducibility was checked in 146 meta-analyses from 82 different papers. As mentioned above, in 5 of these meta-analyses (3%), all from the same published article, the original script code was available. Therefore, in these five cases, outcome reproducibility was checked running the original analysis script on the original primary data. In the remaining cases, the original analytical framework was reconstructed as explained in the method section. Figure 4 summarises the results of the whole process of outcome reproducibility assessment.

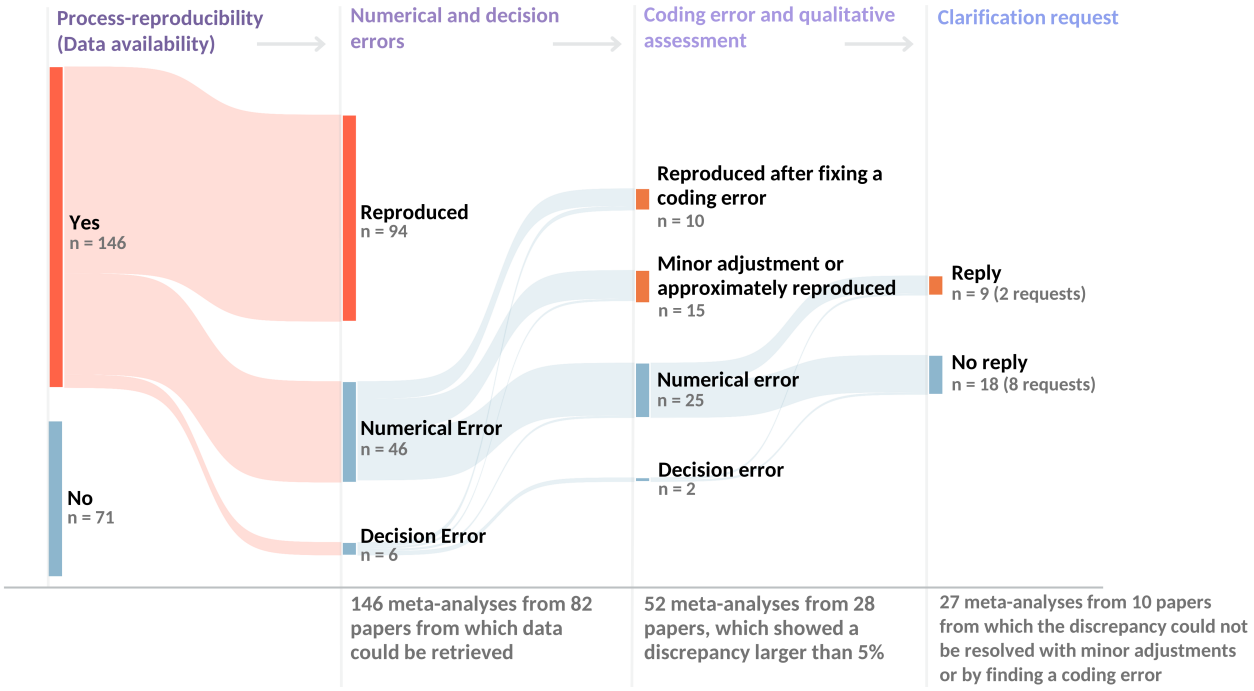

Figure 4

Results of the different stages carried out in the evaluation of the outcome reproducibility.

Following the first stage of re-analysis, 52 meta-analyses were re-assessed because they were labelled as numerical error and/or decision error.

Of these, 17 were re-analysed again as some coding errors were found in the second stage. After this, 10 were re-labelled as reproduced and 7 still had relevant discrepancies.

Furthermore, 15 were labelled as approximately reproduced or reproduced with minor adjustment in a qualitative check because the discrepancy was probably explained by rounding issues, inverted signs for results (when effect sizes were reported in absolute values) and primary data, minor reporting errors, or minor adjustments in the analytical scheme[6]. In the remaining 20, and in the 7 re-analysed again without success, some issues or relevant discrepancies without apparent explanation were found. Figure 5 displays a scatterplot showing the consistency between the original and reproduced summary effect size and their confidence bounds of these 52 meta-analyses. Additionally, as a secondary analysis, the reproducibility of the I^2 heterogeneity statistic was explored. Figure 6 displays a scatterplot showing the consistency between the original and reproduced I^2 statistics. As shown in Figures 5 and 6, the discrepancies found in the heterogeneity statistic I^2 are larger than those found in the summary effects and their confidence intervals. The Pearson's correlation between the summary effect and I^2 discrepancies was .172. The lack of precision of the available data (rounded data) or incomplete information on aspects such as the tau-squared estimator applied seem to have a substantial impact on the reproducibility of this result.

Main issues identified

Different issues in these 27 meta-analyses were identified in the second stage. For example, for one of the meta-analyses which showed a discrepancy in the confidence limits, inconsistencies were found in the original meta-analytic report itself. The confidence limits originally reported for that meta-analysis were different in the abstract, main text and forest plot. Matching the reproduced results were those reported in the forest plot but not those reported in the text. Furthermore, inconsistencies in the original summary effect reported were found between the results reported in abstract and the results reported in the main text and the forest plot. Also, in a paper where primary data were available in both a table and a forest plot, minor inconsistencies were found between the primary data of the table and the

[6] Full details in Supplementary File at: https://osf.io/fjhpw

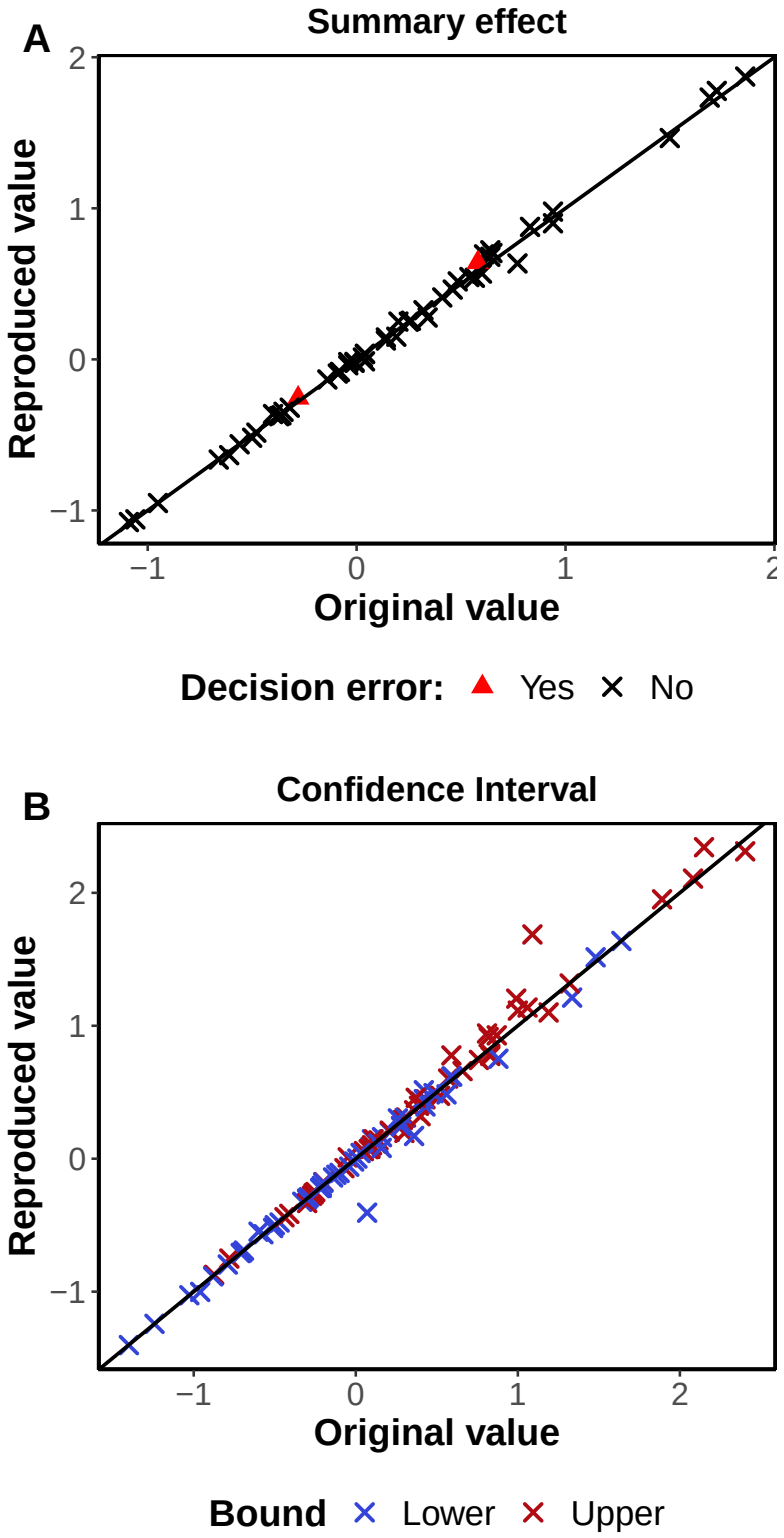

Figure 5

Scatterplot displaying the reproduced values as a function of the original values classified by whether or not decision error was found. Only the results of the 52 meta-analyses with a discrepancy of more than 5% identified in the first stage are displayed, but with the corrections made in the second stage. In panel (a) the summary effects are displayed and in panel (b) the confidence intervals. For (b) the colours represent lower or upper bound of the confidence interval.

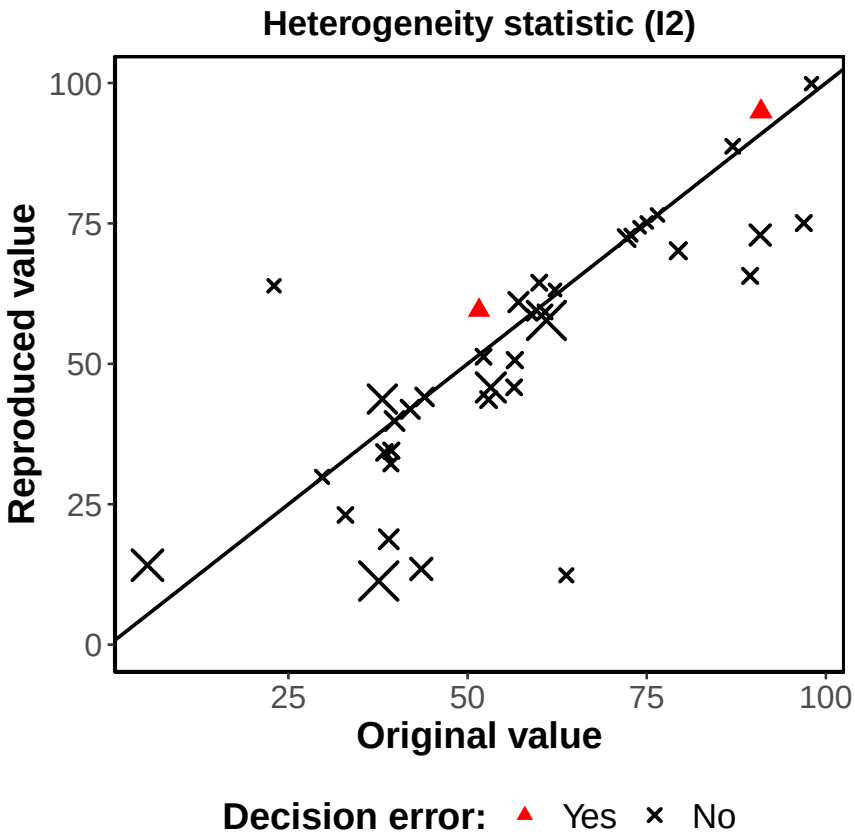

Figure 6

Scatterplot displaying the reproduced values as a function of the original values classified by whether or not decision error was found. Only the results of the 52 meta-analyses with a discrepancy of more than 5% identified in the first stage are displayed, but with the corrections made in the second stage. The values displayed are I^2 heterogeneity statistics. The size of the crosses is a function of the discrepancy in the summary effect.

forest plot. These examples of inconsistencies in original results or data were found in 4 cases (3%, 4/146). These appeared to be typos. Furthermore, some inconsistencies were found with respect to the number of primary studies included in each meta-analysis. For example, in one of the meta-analyses, the main text reported the inclusion of 10 comparisons in the meta-analysis, whereas in a table of results 11 comparisons were reported for this meta-analysis. On the other hand, in 11 meta-analyses the primary data retrieved from the supplementary materials were not sufficient to reach the number of primary studies stated as included in this meta-analysis in the original report.

Original authors clarifications

These 27 meta-analyses were from 10 different papers. Therefore, 10 clarification requests with information about the study aims, methods, and preliminary results were sent to the corresponding authors of the original articles. A reply was received in only 2 of the 10 cases. In one of them, the original authors sent back a link to an OSF repository[7] where the original data and analysis script were stored. According to the authors, this link was not reported in the paper by mistake. The script was run on these data and the results were successfully reproduced. In this case, the data previously used were retrieved from a forest plot (means and standard deviations) and a table (sample sizes) reported in the paper. The previous discrepancy was explained by two cases included in the original meta-analysis from the same primary study that were reported with the same ID in the forest plot and were not correctly matched with their corresponding sample size extracted from the table. This situation exemplifies the potential issues arising from having to reconstruct the original data from tables and figures and not having open access to the original data file.

In the other case, the original data was retrieved from a huge table in supplementary material with all effect sizes and their confidence limits. The original authors sent back this same table by increasing the number of decimal places of the effect sizes and after correcting some wrong values that they themselves detected in that process. This fixed the discrepancies for some of the meta-analyses in this paper.

Discussion

The main aim of this study was to examine the reproducibility of a sample of published meta-analyses on the effectiveness of clinical psychology interventions. We

[7] According to the repository timeline the project was created on 02/06/2019 and according to the journal's article history the paper was published on 13/06/2019. It seems that the repository was created as a journal requirement

analyzed the availability and reusability of original data and, assessed the reproducibility of the published results using these retrieved original data, and tried to reconstruct the original analysis plan. We encountered both difficulties in retrieving the original data and some problems with the reproducibility of the meta-analyses examined.

Even when we interpret data availability in the broad sense (i.e. retrieving data from tables and figures when no data file was available), for about a third of the included meta-analyses no data were available. In these cases, attempts were made to obtain the data on request to the corresponding author, with little success. Authors only shared data in 12% of the requests that were made. This result is in line with what was found in a recent study where data availability statements from a set of primary studies were analysed (Gabelica et al., 2022). Although 42% of primary studies in Gabelica et al. (2022) reported data were available on request (an identical percentage was found in Page et al. (2022) for meta-analyses), only 6.8% of the authors shared the underlying data when requested. Even though it is common to see authors state data is available on request, actually obtaining the data on request seems highly challenging. Although this problem of retrieving data on request is well known (Wicherts et al., 2006), the situation does not seem to have improved. Nowadays, there are straightforward, free, and open ways to share data, including meta-analytic data files. Several repositories (e.g., OSF, GitHub, Zenodo, Figshare) are available for researchers to openly share the data associated with published results. On-request availability has proven to be inadequate, and with the availability of data repositories it is no longer necessary. Journals publishing meta-analyses should require that authors share the underlying data in a public data repository.

Nevertheless, a more positive sign comes from the positive association between publication year and the possibility of retrieving the data. The results tentatively suggest a trend of improving data availability over the years, with a notable rate of 80% observed in meta-analyses published between 2016 and 2020. This observation could be related to the existence of well-established meta-analysis reporting guidelines. For instance, the first

PRISMA guideline (Moher et al., 2009) encouraged meta-analyst to report results of primary studies (e.g. primary effect sizes and their confidence interval through a forest plot, as was a common scenario among the cases included in this project), and the latest PRISMA guideline (Page et al., 2021) which puts more emphasis on appropriate data sharing through data files ready for reuse. At the same time in only 5% of the cases where data were retrieved in our sample were we able to retrieve the data in a machine-readable data file that was ready for reuse (e.g., *csv*, *xlsx*). Most often the data had to be retrieved from files in document format (e.g., *docx*, *pdf*). This forces people who want to reuse the data to manually recode the data, which is an inefficient and error-prone task. Even after partial double coding was carried out, this procedure did not avoid some coding errors, which were only detected by double checking meta-analyses with discrepancies. In our experience, the data retrieval process can be difficult when results are presented in general tables, as it involves matching subsets of these primary data with different meta-analytic results, while is not always clear which studies were used in which meta-analysis reported in a paper. Furthermore, because the tables in manuscript are often generated manually in document file formats (e.g. Word), we observed examples where this introduced another source of error. The foregoing discussion raises a key point about how time consuming the appraisal of meta-analytic reproducibility currently is, and how efficiency would be improved by having open access to the underlying meta-analytic data in data file formats ready for reuse. The latest PRISMA guidelines, along with some initiatives that promote appropriate data sharing (e.g. Wilkinson et al., 2016) have the potential to generate significant improvements in the re-use of meta-analytic data in the years ahead. In this regard, our results provide a useful baseline for future assessments.

An important finding is that the availability of the original analysis script was very limited. Only in five meta-analyses (3%, all from the same paper), was the original script openly available. In most cases, the original analyses were reconstructed from the description provided in the paper itself, which was not always rich in detail, so many of these computational details had to be inferred from the default settings of the software authors

used. The availability of analysis scripts often shows similar rates, both in meta-analyses (Page et al., 2022; Polanin et al., 2020) and in primary research (Hardwicke et al., 2020, 2022). This makes it more difficult to easily check the computational reproducibility of the results from such studies. Reconstructing the analytical scheme adds to the workload, with the potential to introduce errors, both in the original report and in the reconstruction, and deals with the eventual lack of relevant analytical information. With the increasing availability of excellent open-source tools to perform meta-analysis (e.g., *metafor* (Viechtbauer, 2010) in R) and useful templates (Moreau & Gamble, 2020), meta-analysts can use workflows that allow them to create and share analysis code for meta-analyses.

Despite these difficulties, we were able to recover the original data and reconstruct the original analysis approach, for 146 meta-analyses, for which the reproducibility of the results was assessed. These attempts went through several stages as explained above, trying to minimize the impact of possible coding errors, and requesting clarifications from the original authors. Nevertheless, even with these efforts, some discrepancies remained in the results. We identified different issues that hindered our reproducibility attempts. For example, in some cases internal discrepancies were found in the paper itself (e.g., text-figure discrepancies, text-abstract, or text-table discrepancies). Furthermore, some problems were found with the lack of some primary data, where data available in the supplementary material included fewer cases than those finally reported in the results of the published paper. These situations could be explained by typos in the manuscript, or updates when performing the meta-analysis that produced different versions of the manuscript, data, or supplementary material. While it is important to note that discrepancies in the summary effect results and their confidence intervals were mostly minor, with little or no impact on the conclusions, these situations are easily avoidable. Some of the problems identified could be explained by typos. Currently, there are tools that facilitate the production of so-called reproducible manuscripts, such as the R packages *knitr* (Xie, 2022), *rmarkdown* (Allaire et al., 2022), and *papaja* (Aust & Barth, 2022). A reproducible manuscript embeds analysis code, data and

results reporting in a single document, extracting and reporting the results from the output of the computational process itself, avoiding error-prone manual transcriptions.

Our results are complementary to those observed in previous research on the reproducibility of the primary effects of meta-analyses (Gøtzsche et al., 2007; Maassen et al., 2020) and related problems due to the multiplicity of primary effects (Tendal et al., 2009). These studies found problems in reproducing the primary effects of published meta-analyses, or in reaching agreement between independent coders in computing them. Such problems, to a greater or lesser extent, had some impact on the meta-analytic results. Our results show that, even when re-using the primary effects as originally coded, certain problems of reproducibility of the results may remain. Some of these problems are added error on the source of error found in previous research on reproducibility of primary effects, which in turn are added error on the sources of error types of primary estimates (e.g., measurement error, sampling error, or reporting errors). No scientific research is totally error-free, but one of the main tasks of scientists is to minimize this error, and in some cases, such as those observed in this study, minimizing some potential sources of error can be straightforward.

Our study has some limitations. First, the time span covered is fairly wide. Thus, the findings may not capture the changes that have arisen in recent years. Therefore, future studies should examine more specific changes over years, to evaluate whether better practices emerge that facilitate reproducibility. Second, most of the primary data was retrieved through manual re-coding, which introduces some error. The reported data was rounded, which means we did not have access to precise values, and in many cases the standard error had to be approximated from the confidence limits. These limitation of our study are caused by the suboptimal practices when sharing data we discussed above. Given the non-precise nature of most of the data retrieved, we had to make a decision about which margin of discrepancy was acceptable. In this study, a margin of 5% was chosen. Because this cut-off is arbitrary, we have tried to focus more on possible issues in the results that fell above this margin, than on establishing a exact ratio of non-reproduced meta-analyses based on this

arbitrary cut-off. Finally, we only examined meta-analyses in clinical psychology as this is one of the areas that produces the most meta-analyses in psychology and these meta-analyses have a direct impact on applied practice, but it is unknown to which extent our conclusions generalize to meta-analyses in other sub-disciplines in psychology.

In conclusion, we observed several difficulties when attempting to reproduce meta-analyses. Two aspects can be highlighted: (1) data availability and reusability of the data as they are shared, (2) and apparent errors in the reporting of results. As data collected for a meta-analysis can be especially useful for future research, direct and open access to such datasets allows for easy updates, and re-analyses, which are valuable in a cumulative science. Meta-analytic data generally do not contain sensitive or personal information, and can therefore almost always be shared openly, as doing so does not involve ethical or legal conflicts. Third, meta-analytic results often represent the state of the art of the evidence on a particular topic. These results guide applied practice, public policy, or future research directions. This prominent status entails a major responsibility for the credibility, reliability, and validity of published meta-analytic results.

Author Contributions

Conceptualization: R. Lopez-Nicolas and J. Sanchez-Meca; Methodology: R. Lopez-Nicolas, D. Lakens, J.A. Lopez-Lopez and J. Sanchez-Meca; Formal Analysis: R Lopez-Nicolas; Investigation: R Lopez-Nicolas, M. Rubio-Aparicio, A. Sandoval-Lentisco, C. Lopez-Ibañez and D. Blazquez-Rincon: Data curation: R Lopez-Nicolas; Writing – Original Draft Preparation: R. Lopez-Nicolas; Writing – Review & Editing: D. Lakens, J.A. Lopez-Lopez, J. Sanchez-Meca, M. Rubio-Aparicio, A. Sandoval-Lentisco, C. Lopez-Ibañez, D. Blazquez-Rincon.

Conflict of interest

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments

We thank the research group of the Meta-Research Centre of Tilburg University for providing feedback on an earlier version of this manuscript. We would also like to thank all the original authors of the included meta-analyses who enabled access to the data used in this project as well as the authors who provided clarifications in response to requests for clarification in the third stage.

Funding

This research has been funded with a grant from the Spanish Ministry of Science and Innovation (Project PID2019-104080GB-I00/AEI/10.13039/501100011033, and PID2019-104033GA-I00/MCIN/AEI/10.13039/501100011033, FEDER funds) and by the Spanish Ministry of Universities (predoctoral grant: FPU18/04805).

Supplemental Material

Supplementary material available at: https://osf.io/fjhpw

Prior versions

A version of this manuscript has been posted as a preprint on PsyArxiv: https://psyarxiv.com/gvqrn/

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). *Rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/met0000365

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 1–15. https://doi.org/10.1057/s41599-021-00903-w

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2023). *Rcrossref: Client for various 'CrossRef' 'APIs'*.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: Mixed-methods study. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2022.05.019

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

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175–182.

https://doi.org/10.1038/nature25753

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society Open Science*, *8*(1), 201494. https://doi.org/10.1098/rsos.201494

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). *Royal Society Open Science*, *7*(2), 190806. https://doi.org/10.1098/rsos.190806

Koffel, J. B., & Rethlefsen, M. L. (2016). Reproducibility of Search Strategies Is Poor in Systematic Reviews Published in High-Impact Pediatrics, Cardiology and Surgery Journals: A Cross-Sectional Study. *PLOS ONE*, *11*(9), e0163309. https://doi.org/10.1371/journal.pone.0163309

López-Nicolás, R., López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2022). A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavior Research Methods*, *54*(1), 334–349. https://doi.org/10.3758/s13428-021-01644-z

Maassen, E., Assen, M. A. L. M. van, Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Maggio, L. A., Tannery, N. H., & Kanter, S. L. (2011). Reproducibility of Literature Search Reporting in Medical Education Reviews. *Academic Medicine*, *86*(8), 1049–1054. https://doi.org/10.1097/ACM.0b013e31822221e7

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*, *27*(3), 426–432. https://doi.org/10.1037/met0000351

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. The National Academies Press. https://doi.org/10.17226/25303

Nguyen, P.-Y., Kanukula, R., McKenzie, J. E., Alqaidoom, Z., Brennan, S. E., Haddaway, N. R., Hamilton, D. G., Karunananthan, S., McDonald, S., Moher, D., Nakagawa, S., Nunan, D., Tugwell, P., Welch, V. A., & Page, M. J. (2022). *Changing patterns in reporting and sharing of review data in systematic reviews with meta-analysis of the effects of interventions: A meta-research study*. medRxiv. https://doi.org/10.1101/2022.04.11.22273688

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

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

Page, M. J., Altman, D. G., Shamseer, L., McKenzie, J. E., Ahmadzai, N., Wolfe, D., Yazdi, F., Catalá-López, F., Tricco, A. C., & Moher, D. (2018). Reproducible research practices are underused in systematic reviews of biomedical interventions. *Journal of Clinical Epidemiology*, *94*, 8–18. https://doi.org/10.1016/j.jclinepi.2017.10.017

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Page, M. J., Nguyen, P.-Y., Hamilton, D. G., Haddaway, N. R., Kanukula, R., Moher, D., & McKenzie, J. E. (2022). Data and code availability statements in systematic reviews of interventions were often missing or inaccurate: A content analysis. *Journal of Clinical Epidemiology*, *147*, 1–10. https://doi.org/10.1016/j.jclinepi.2022.03.003

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., Catalá-López, F., Li, L., Reid, E. K., Sarkis-Onofre, R., & Moher, D. (2016). Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLOS Medicine*, *13*(5), e1002028. https://doi.org/10.1371/journal.pmed.1002028

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and Reproducibility of Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1),

192. https://doi.org/10.1038/s41597-021-00981-0

Tendal, B., Higgins, J. P. T., Jüni, P., Hróbjartsson, A., Trelle, S., Nüesch, E., Wandel, S., Jørgensen, A. W., Gesser, K., Ilsøe-Kristensen, S., & Gøtzsche, P. C. (2009). Disagreements in meta-analyses using outcomes measured on continuous or rating scales: Observer agreement study. *BMJ*, *339*, b3128. https://doi.org/10.1136/bmj.b3128

Tendal, B., Nüesch, E., Higgins, J. P. T., Jüni, P., & Gøtzsche, P. C. (2011). Multiplicity of data in trial reports and the reliability of meta-analyses: Empirical study. *BMJ*, *343*, d4829. https://doi.org/10.1136/bmj.d4829

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Wallach, J. D., Boyack, K. W., & Ioannidis, J. P. A. (2018). Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biology*, *16*(11), e2006930. https://doi.org/10.1371/journal.pbio.2006930

Wayant, C., Page, M. J., & Vassar, M. (2019). Evaluation of Reproducible Research Practices in Oncology Systematic Reviews With Meta-analyses Referenced by National Comprehensive Cancer Network Guidelines. *JAMA Oncology*, *5*(11), 1550–1555. https://doi.org/10.1001/jamaoncol.2019.2564

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726. https://doi.org/10.1037/0003-066X.61.7.726

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Xie, Y. (2022). *Knitr: A general-purpose package for dynamic report generation in r.* https://yihui.org/knitr/