UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Research Synthesis: Challenges, Reproducibility and Future Directions.

Síntesis de la Investigación: Desafíos, Reproducibilidad y Direcciones Futuras

D. Rubén López Nicolás

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Research Synthesis: Challenges, Reproducibility and Future Directions

Síntesis de la Investigación: Desafíos, Reproducibilidad y Direcciones Futuras

Autor: D. Rubén López Nicolás

Director/es: Dr. Julio Sánchez Meca y Dr. José Antonio López López

**DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO DE DOCTOR**

Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Rubén López Nicolás

doctorando del Programa de Doctorado en

Psicología

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Research Synthesis: Challenges, Reproducibility and Future Directions/Síntesis de la Investigación: Desafíos, Reproducibilidad y Direcciones Futuras

y dirigida por,

D./Dña. Julio Sánchez Meca
D./Dña. José Antonio López López
D./Dña.

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones o, tal y como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 4 de septiembre de 2023

Fdo.:

If I have seen further, it is by standing on the shoulders of giants
(Newton to Robert Hooke, 1675)

Agradecimientos

El camino que me ha llevado hasta este punto, el punto en el que me encuentro dando las últimas pinceladas a lo que será mi Tesis Doctoral, ha sido un camino largo y lleno de personas que merecen su mención en este apartado, que merecen su agradecimiento por haber hecho esto posible. Espero no olvidarme de nadie y tener un hueco aquí para todos y todas los que lo merecen.

En primer lugar, a mi familia. A mi madre y mis hermanas por inspirarme y empujarme a crecer y no conformarme. Y a mi padre, que sé que me está mirando con orgullo allá donde este. Soy lo que soy y estoy donde estoy por vosotras. También me gustaría mencionar a Adrián y Alberto, mis dos cuñados, los cuales son parte fundamental de mi familia.

A mis compañeros y compañeras de carrera, en especial a Ismael, Isa, Claudia y Fran. Fuisteis y sois fuente de motivación constante y tengo completamente claro que sin haberme cruzado con vosotros en este momento no estaría a un paso de convertirme en Doctor. Vaya cuatro años nos pasamos, de las mejores épocas de mi vida.

A mis amigos. Los de siempre, Iván, Sergio, Jesús, Javi, Alex y Salva, vosotros que me habéis visto en muchas circunstancias estaréis igual de alucinados que yo de que la vida me haya traído hasta este punto. Los que han estado más cerca en estos últimos años, José, Juan Antonio, Dani, Pablo^2 y Diego, habéis sido una fuente fundamental apoyo durante todo este proceso, sin vosotros todo habría sido mucho más difícil. Y a los que conocí como compañeros de unos de esos trabajos que me ayudó a poder terminar la carrera que hoy me hace estar escribiendo esto y desde entonces siempre han estado cerca, José, Luis y Dani.

Por su puesto, a mis compañeros de fatigas durante estos cuatro años. A los predoctorales de Psicología Básica y Metodología, Lucia, Víctor, Carmen, Desirée, Alex, Míriam y a nuestra reciente incorporación, Ana. Que decir de lo que habéis sido, de lo que hemos sido. Haber transitado todo este camino a vuestro lado lo ha hecho todo muchísimo más fácil. Hemos pasado de todo, nos hemos ayudado, aconsejado, apoyado y hecho mejores los unos a los otros. Ha sido un placer haber compartido todo esto con vosotros.

A mi grupo de investigación, el cual me acogió desde el principio, cuando aún era un simple estudiante de grado interesado por la investigación, que quiso dar sus primeros pasos como alumno interno. A María, que desde el principio fue una importantísima fuente de apoyo y consejo. A Fulgencio, una persona que inspira en todo lo que hace. Y por supuesto a mis directores de tesis Julio y José, gracias por aceptarme como doctorando y por hacer posible que hoy esté escribiendo esto, por hacer posible que, en Murcia, mi ciudad, haya un equipo y entorno de investigación puntero sin el cual nada de esto habría sido posible.

A todos los colaboradores que han participado en los trabajos que conforman esta tesis doctoral, muchos de los cuales ya han sido mencionados, pero quiero hacer especial mención a mis dos supervisores de las estancias que he realizado durante mi formación predoctoral. Daniël Lakens, que me acogió bajo su supervisión en la Universidad Tecnológica de Eindhoven y con el cual ha sido un placer trabajar y colaborar. Fueron tres meses fantásticos en Países Bajos. Tan buenos fueron que repetí país en mi segunda estancia, esta vez en la Universidad de Tilburg, bajo la supervisión de Robbie van Aert, que me acogió, junto a todo su grupo, y con el cual ha sido un placer trabajar y colaborar. Muchas gracias a los dos.

Por último, mencionar a dos personas que han estado muy cerca de mí durante estos cuatro años, en momentos distintos, pero con aportes fundamentales en apoyo y motivación. Gracias, Lucia y Alicia.

Title

Research Synthesis: Challenges, Reproducibility and Future Directions.

Abstract

Research synthesis projects play an indispensable role in the scientific process as they bring order to the vast array of scientific evidence, organizing individual pieces of evidence into a coherent body of knowledge on a specific topic. Given this prominent role, the results and conclusions of research synthesis projects carry greater relevance and impact compared to those of individual studies. Therefore, it is essential to keep an eye on the research practices and credibility of research synthesis projects. In this dissertation, we delve into various aspects of research practices and the credibility of research synthesis projects. The first study (Chapter 2) focused on assessing the prevalence of transparency and reproducibility-related reporting practices in research synthesis projects. The second study (Chapter 3) focused on reproducibility of meta-analytic results reported on these projects. Lastly, the third study (Chapter 4) explored the statistical power of meta-analytic synthesis when assuming a random-effects model. All three studies were carried out using a random sample of 100 published research synthesis projects on effectiveness of clinical psychological interventions.

Keywords

Research synthesis; meta-analysis; systematic review; meta-science; reproducibility; data sharing.

CONTENTS

Chapter 1

General Introduction

Scientific progress relies on the accumulation of knowledge by building upon the prior contributions of fellow researchers. Commonly, no single study provides enough information to answer a question conclusively. Low precision, low statistical power, and the specificity of a particular context and design threaten the reliability of the estimates, the probability of detecting true effects, and the generalizability of the conclusions drawn in single studies.

Furthermore, the number of available scientific publications is constantly growing, with the growth rate escalating each year (Bornmann et al., 2021). Therefore, scientific literature abounds with singles studies –many of them focusing on the same or similar phenomena— which provide partial information that should be synthesized in order to reach more reliable and useful answers to scientific questions. As early as the late 20th century, Morton Hunt (1997) eloquently articulated this notion, stating:

> Fundamental assumption that our culture makes about science, namely, that is progressive and cumulative…for centuries it has been an article of faith that scientists base their research on existing information, add a modicum of new and better data to it, and thereby advance toward an ever more profound, complete, and accurate explanation of reality.
>
> But today we are experiencing a crisis of faith; many of us no longer feel sure that science, though. Growing explosively, is moving inexorably toward the truth. Indeed, "growing explosively" is an ominous oxymoron: growing implies orderly development, but explosively denotes disorder and fragmentation. (p. 1)

In this context, research synthesis approaches have become an essential part of scientific progress. However, it is not an issue of whether to synthesise or not, it is more about how to synthesise. Different approaches have been used throughout history to synthesise the available research. The strength and limitations of each have led to those most widely currently used. Next, a brief overview of the different approaches used at different times in history is presented.

1.1 Research synthesis over time

1.1.1 From narrative reviews to systematic reviews

One of the earliest forms of research synthesis was the so-called narrative reviews. This kind of reviews were carried out by experts on a particular topic who provided a broad overview of a specific research area or question. Narrative reviews are often conducted in an unsystematic and subjective manner which introduces a high risk of bias in the conclusions drawn from such reviews. According to White (2019):

> Narrative reviewers had been mute on how they found the studies under review…had accepted or rejected studies impressionistically…were inconsistent in deciding which aspects of studies to discuss…used ad hoc judgments as to the meaning of statistical findings. (p. 53)

These methodological shortcomings were pointed out at the end of 20th century by the social scientist David Pillemer (1984):

> Meta-analysts characterize the usual review as subjective, relying on idiosyncratic judgments about such key issues as which studies to include and how to draw overall conclusions. Studies are considered one at a time, with strengths and weaknesses selectively identified and casually discussed. Since the process is informal, it is not surprising that different reviewers often draw very different conclusions from the same set of studies. Left to confront tens or even hundreds of studies without formal tools, the narrative reviewer must rely on personal strategies to coax out reliable findings. (p. 28)

Consequently, narrative reviews were often considered as scientifically unsound publications.

As the amount of evidence available increased, the relevance of research synthesis also increased, leading to the urge to treat the process of research synthesis as a scientific enterprise (Chalmers et al., 2002). Recognizing the limitations and subjectivity of narrative reviews, researchers sought to develop more rigorous methods that would provide a systematic and transparent approach to synthesizing research findings. This marked a significant shift in the field of research synthesis, prompting the development of new approaches that aimed to minimize bias and enhance the reliability of the synthesized evidence. Gregg B. Jackson (1980) and Harris M. Cooper (1982) laid the groundwork for research synthesis as a research process, drawing an analogy between research synthesis and primary research, describing the different stages of the process, their function, and their potential threats to validity. These endeavours materialised with the advent of so-called systematic reviews.

1.1.1.2 Systematic reviews

Systematic reviews are characterized by focusing on well-defined research questions, seeking to shape a proper design for those aims; by their comprehensiveness, seeking to capture all the available evidence on a specific topic; by predefined inclusion and exclusion criteria, seeking to avoid selection biases; by carrying out critical appraisal of included studies, seeking to measure the validity of the evidence considered; by using predefined approaches to synthetise the results of included studies – both quantitative and qualitative, seeking to avoid potential biases.

As an astute reader may have noticed, systematic reviews are characterized by a systematization of review processes. Nonetheless, what truly establishes systematic reviews as a scientific endeavour is carrying out and reporting them in a transparent and reproducible manner.

1.1.2 Quantitative synthesis

Another key issue on how to properly conduct an accurate research synthesis is how to draw sound conclusions from a collection of individual studies. Traditionally, narrative reviewers addressed this matter following a sort of quasi-quantitative proto-vote-counting approach. Basically, from the collection of reviewed studies, narrative reviewers drew conclusions discussing the results of the set of individual studies eventually reaching a sort of consensus on the majority direction. Although the vote-counting approach underwent systematization and refinement (Hedges & Olkin, 1980;

Light & Smith, 1971), nowadays its use is not recommended in most circumstances or even deemed an unacceptable synthesis method (McKenzie and Brennan, 2022).

On the other hand, since Karl Pearson (1900) described the χ^2 distribution and used the χ^2 statistic to test the independence of proportions, and later Ronald A. Fisher (1925) formalized the concept and broadened its applicability to researchers, *p-values* have been ubiquitous in quantitative research. Naturally, there were proposals to use them in quantitative synthesis. Fisher (1932) and Pearson (Pearson, 1938) themselves, along with others (e.g., Stouffer et al., 1949; Wilkinson, 1951), made proposals for the combination of independent *p-values* from individual studies. While these methods can be useful in certain scenarios or for specific purposes (McKenzie and Brennan, 2022), they do not address the most informative aims when quantitative synthesis is carried out (Hedges & Olkin, 1985), namely estimating the average magnitude of the effects and assessing consistency across individual studies. Bearing these two aspects in mind, the following section briefly summarizes the history of what is currently the predominant approach to quantitative synthesis, the so-called meta-analysis.

1.1.2.1 Meta-analysis

Karl Pearson's (1904) work on effectiveness of a vaccine against typhoid fever could be considered as the earliest meta-analysis in its contemporary sense. He computed correlation coefficients for a set of 11 individual studies, pooled them within two subgroups and discussed the observed variability among them. Subsequently, William G. Cochran, through a series of works (1937, 1953, 1954) introduced and developed fundamental concepts that continue to be utilized in modern meta-analysis, namely weighting individual effect sizes by precision (e.g., inverse variances) and estimating the heterogeneity among primary effect sizes beyond what can be attributed to their sampling variances.

These prior endeavours eventually catalysed when Gene Glass coined the term *meta-analysis* at the 1976 Annual Meeting of the AERA (American Educational Research Association). Furthermore, a few years later, the influential handbook *Statistical Methods for Meta-Analysis* by Larry V. Hedges and Ingram Olkin was published in 1985, solidifying the field of meta-analysis. Nonetheless, it is worth mentioning that these advancements were not immune to criticism. Scholars such as Eysenck (1978) and Shapiro (1994) raised critical viewpoints regarding certain aspects of meta-analysis.

1.2 Research synthesis today

Nowadays, systematic reviews and meta-analyses have gained widespread recognition as the gold standard approach for research synthesis. Numerous up-to-date handbooks provide comprehensive guidance (e.g., Cooper et al., 2019; Higgins et al., 2019; Schmid et al., 2021) on conducting these syntheses, offering detailed instructions and recommendations. Moreover, guidelines for proper reporting of syntheses (e.g., Page et al., 2021; Sánchez-Meca et al., 2021) outline the essential elements that should be included when reporting on research synthesis work.

Although this field is dynamic and constantly evolving, with continuous development of new methods and refinement of existing ones, it has established a solid foundation. This enables the production of informative and reliable work, allowing researchers to generate valuable insights and evidence-based conclusions. Hence, systematic reviews and meta-analysis are frequently considered as top-tier sources of scientific evidence in the context of evidence-based practice in psychology and related fields.

The subsequent section outlines a well-established, multi-stage pipeline that delineates research synthesis as a scientific process. This pipeline draws upon the work of Harris M. Cooper (1982; 2017; 2019) and provides a comprehensive framework for conducting research synthesis.

1.2.1 Research synthesis as a scientific process

1.2.1.1 Formulating the problem

As in any scientific endeavour, the first step in a systematic review is formulating the problem. This stage requires clearly defining the research question to be addressed in the synthesis. The correct formulation of the research question involves establishing both the operational and conceptual definitions of the constructs and variables involved, as well as defining the relationship between these variables. These well-defined formulations serve as a guide throughout the entire synthesis process. They inform various aspects of the review, such as determining eligibility criteria for including studies, conducting systematic searches to identify relevant studies, collecting data from the included studies, structuring the synthesis process, and presenting the findings.

1.2.1.1.1 Eligibility Criteria

Once the research question and variables of interest have been clearly defined, the next step in a systematic review is to pre-specify the inclusion criteria (McKenzie et al., 2022). These criteria play a crucial role in determining which studies will be included in the synthesis and are therefore of utmost importance. The eligibility criteria cover various aspects, including the population of interest, the study design(s) that will be included, the types of interventions or exposures under investigation, the outcomes of interest, years covered by the synthesis, publication status of the included studies, and any other relevant characteristics of the studies. The eligibility criteria serve as a set of predefined rules that guide the study selection process, ensuring that only studies meeting these criteria will be considered for inclusion in the synthesis.

1.2.1.2 Searching the literature

The next step in the systematic review process involves conducting a comprehensive search to identify primary research studies that have the potential to be included in the synthesis. Unlike primary research, which typically targets a finite sample of units (e.g., participants), systematic reviews aim to include all empirical studies on a specific topic of interest that meets the predefined inclusion criteria. This means that the search must be exhaustive, aiming to identify as many eligible studies as possible. Therefore, is recommended to use a range of sources such as including bibliographic databases, journals, conference proceedings, grey literature, and expert recommendations.

In order to conduct a successful search in a systematic review, it is necessary to design a proper search strategy taking into account its comprehensiveness and its precision. Comprehensiveness or sensitivity of a literature search refers to the ratio of relevant studies retrieved by the search strategy to all the relevant studies available in the literature. On the other hand, precision or specificity refers to the ratio of relevant studies retrieved by the search strategy to all the retrieved studies. These two elements are inversely related to each other. Despite, more comprehensiveness implies more workload, systematic reviews aim to include all relevant studies that meet the predefined inclusion criteria to ensure a representative synthesis of the available evidence. Therefore, it is generally recommended to prioritize comprehensiveness over precision.

1.2.1.2.1 Selecting studies

Once the search process is completed, and considering the emphasis on comprehensiveness over precision, it is expected that the output set will contain a certain degree of noise or irrelevant studies. To identify and select the studies that meet the predefined inclusion criteria, a selection process is required, which often involves multiple stages. First, the most obviously irrelevant studies can be identified and removed based on a preliminary assessment of their titles and abstracts. In this stage, all potentially relevant studies are retained. Then, the full-text reports of those studies are examined and faced with the predefined inclusion criteria to determine their eligibility for inclusion in the synthesis. The decisions made at this stage are among the most influential decisions that are made in the synthesis process, so it is highly recommended to employ a double independent screening approach.

1.2.1.3 Data collection and critical appraisal

The selected primary studies typically serve as the observations or units of analysis for the synthesis. Therefore, it is necessary to extract information from them. During the data collection process, several variables such as primary or secondary outcomes, potential moderators, study characteristic, methodological details, and other relevant factors are coded from the included studies. The specific information to be extracted and coded should align with the research question and objectives of the synthesis.

Another relevant source of information from primary studies is study quality. It can be seen as a measure of the internal validity of the results gathered from a primary study. Considering study quality becomes essential when assessing and synthetising a collection of studies that, among other factors, will vary in study quality. Study quality indicators allow us to set them as inclusion/exclusion criteria, explore relationships between them and other characteristics of the studies, or test if effect sizes vary as a function of study quality (Valentine, 2019).

1.2.1.4 Synthetising studies

The synthesis of the collection of included studies, utilizing the extracted information, represents the core stage of the process. It is during this stage that we obtain results leading to meaningful insights. While various approaches exist for conducting the synthesis process, our focus lies on the most widely used approach, meta-analysis.

1.2.1.4.1 Meta-analysis

Meta-analysis has been defined as the "use of statistical methods to combine data across studies in order to estimate parameters of interest" (Schmid et al., 2021, p. 19) or "statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings (Glass, 1976).

A standard meta-analysis of aggregate data involves a two-stage process. First, an effect size (by effect size, we refer to the magnitude of a phenomenon's presence in the population) in a common metric for each included study is computed from summary data previously collected. Secondly, a weighted average is computed by combining the primary effect sizes, and the heterogeneity among these effect sizes is estimated. These computations serve the following aims (Schmid et al., 2021):

- Increase the precision of the estimates.
- Quantify the consistency among primary findings.
- Explore the causes of observed inconsistency.

As the aims of meta-analysis show, average estimation, and consistency in the results of the included studies are key points when carrying out a meta-analysis. Therefore, prior to fitting a meta-analytic model, certain assumptions need to be made in this regard. The most common approaches to address these aspects are the fixed-effect and random-effects models.

1.2.1.4.1.1 Fixed and random effect(s) models

Under the standard fixed-effect model, it is assumed that there is a single a common parametric effect for all estimates of the included studies (i.e., $\theta_1 = \cdots = \theta_k = \theta$). Consequently, any variation among study results is assumed to arise solely from sampling error in the primary studies. The purpose of meta-analysis under this model is to improve the precision of the estimation by combining information and to increase the statistical power to detect a true effect. However, in practice, the assumption of a common underlying effect among a collection of studies that differ in various aspects –such as research setting, population of interest, measurement tools among others– is often unrealistic (Higgins et al., 2009).

For this reason, although there are more realistic interpretations for the former approach (Rice, 2018), the most widespread approach in practice is the random-effects model. Under a random-effects model, it is assumed that primary effect sizes represent estimates of a random sample of parametric effects, which are drawn from an underlying distribution, such that $y_i \sim N(\mu_\theta, \tau^2 + \sigma_i^2)$, with y_i being the effect estimate of the ith study, μ_θ denotes the expected value, τ^2 the between-studies variance component, and σ_i^2 the sampling error of the *ith* study. Therefore, the purpose of meta-analysis under this model it to describe this distribution by estimating the expected value of the effect size across the different research settings, populations and any other factors in which the included studies may vary, as well as to assess the consistency of effect sizes between those factors. This model can be readily extended by adding fixed covariates that could account for some of this heterogeneity.

1.2.1.5 Interpreting and reporting

Drawing sound conclusions from synthesis results can be highly context dependent as it relies on the specific topic of interest and the research question at hand. Effect sizes are the common currency of quantitative synthesis and interpreting them is closely tied to the research area (Schäfer & Swartz, 2019) and the specific aims of the synthesis. Furthermore, when interpreting meta-analytic results, it is crucial to consider the strength of the evidence that has been included. This entails taking into account various domains such as the study quality of the included studies, imprecision of the summary estimates of the meta-analytic model, inconsistency among the included studies, and potential reporting biases (Balshem et al., 2011). Reporting biases refer to missing results in the synthesized body of evidence, which can occur due to practices such as outcome selective reporting or publication bias. These practices have the potential to introduce systematic biases that can inflate effect size estimates (Schäfer & Swartz, 2019). There are several methods and approaches available to assess the potential impact of these practices on meta-analytic results. (Carter et al., 2019; Marks-Anglin & Chen, 2020).

Once the synthesis has been completed, it is crucial to report it transparently, with a correct structure, and provide sufficient information so that another researcher can reproduce each stage in the same manner as the original researchers and obtain the same results (Mayo-Wilson and Grant, 2019). There are many reporting guidelines that assist synthesists in this endeavour. One of the most relevant guidelines for reporting meta-analyses is the Preferred Reporting Items for Systematic Reviews and Meta-Analyses

(PRISMA2020) statement (Page et al., 2021). It consists of a set of 27 items that cover the necessary information to be included in the report from the title to the conclusions.

1.3 Credibility crisis

Over the decades, numerous methodologists have drawn attention to various phenomena that could undermine the credibility of published psychological findings. One such phenomenon is the ubiquity of statistically significant results in published literature (Greenwald, 1975; Sterling, 1959), which coexists with a high prevalence of underpowered studies (Chase and Chase, 1976; Cohen, 1962). This paradoxical combination raised different concerns. Rosenthal (1991) drew attention to the "file drawer problem" – which is now commonly referred to as publication bias. It refers to a systematic bias in the publication process that favours statistically significant results.

Despite these early concerns, it was not until the 2010s that some remarkable empirical and conceptual contributions brought about a crisis or revolution in the field of psychology (Nelson et al., 2018; Nosek et al., 2022). The emergence of several replication failures (e.g., Doyen et al., 2012; Galak et al., 2012; Open Science Collaboration, 2015) along with evidence of a troublingly high prevalence of so-called *Questionable Research Practices* (John et al., 2012) –such as *p-hacking* (Simmons et al., 2011) or *HARKing* (Kerr et al., 1998)– shook the credibility of published scientific results.

1.3.1 Meta-science

These facts eventually led to the emergence of an entirely new research field known as meta-science. Meta-science focuses on studying the scientific process itself, including the methods, practices, and incentives that shape research outcomes (Hardwicke et al., 2020).

From this field, a comprehensive framework for the empirical assessment of scientific credibility has been developed (LeBel et al., 2018; Nosek et al., 2022). When assessing the credibility of scientific results, various approaches can be employed. *Reproducibility* involves attempting to obtain the same results as reported in the original publication by using the same data and following the same procedure. It aims to assess whether the findings can be independently verified using the provided information. *Robustness* refers to evaluating the sensitivity of the original results and conclusions to variations in the original procedure while using the same data. This assessment helps determine the stability of the reported findings under different paths of data processing or analysis. *Replicability*, which is a fundamental principle of the scientific method, entails independent researchers addressing the same research question using a similar approach but collecting new data. The aim is to observe consistent results that support the initial findings.

In recent years, several meta-scientific projects have assessed these facets of research credibility (e.g., Hardwicke et al., 2018; Open Science Collaboration, 2015; Silberzahn et al., 2018), revealing some concerns, flaws, and questionable practices within the scientific research process. These findings have raised awareness and promoted discussions about improving scientific practices. As part of these efforts, meta-scientists have contributed with several initiatives or reform proposal to improve the credibility of scientific results. For instance, pre-registration (Wagenmakers et al., 2012) has been proposed as a tool to prevent *p-hacking* and *HARKing*. In essence, pre-registration involves time-stamped reports about hypothesis, design, analysis plan, and other relevant details published before any data are analysed or even collected. Additionally, open data and material sharing have been widely promoted (e.g., Miguel et al., 2014; Wicherts, 2011) as a straightforward way of being able to assess reproducibility of published results, with the aim of detecting potential errors (Bakkers & Wicherts, 2011) or even fraud cases (Simonsohn et al., 2023). In the same vein, open script codes sharing is also a crucial aspect that enables the assessment of computational reproducibility of research project outputs (Kitzes, 2018). Overall, all these proposals revolve around the same idea: moving towards a more transparent workflow in scientific research.

1.3.2 Replicability and reproducibility of research synthesis

The majority of empirical meta-scientific assessments have predominantly focused on primary studies, with comparatively less attention given to evidence synthesis projects. Nonetheless, noteworthy contributions have been made in this area. For instance, Gøtzsche et al. (2007) recomputed the primary effect sizes from 27 meta-analyses by recoding data from primary studies, finding issues in 10 of them. Tendal et al. (2009) recomputed the primary effect sizes and summary meta-analytic estimates by doubly re-extracting the relevant primary statistics by independent coders, finding substantial inconsistencies between coders. In a similar way, Tendal et al. (2011) found that multiplicity of effect sizes in primary studies can lead to different meta-analytic conclusions depending on how such multiplicity is addressed. Lakens et al. (2017) struggled to reproduce a set of meta-analyses due to lack of access to raw data and incomplete reporting of the methodology followed. Kvarven et al. (2020) compared the results of published meta-analyses to large-scale replications on the same topic, finding significant differences in effect sizes for 12 out of the 15 pairs. And last, Maassen et al. (2020) found several challenges in reproducing the calculation of effect sizes based on the information reported by the original authors of each meta-analysis.

The majority of these aforementioned studies focused on the reproducibility of the primary effect sizes, which serve as the unit of analysis in quantitative synthesis or meta-analysis, recomputing them by recoding the data from the primary studies included in the synthesis. However, as discussed in earlier sections, research synthesis as a scientific endeavour is a systematic multi-stage process. Each stage of this process generates specific outputs that can be subject to reproduction attempts. For instance, Koffel & Rethlefsen (2016), assessed the reproducibility-related reporting practices of search strategies in a set of systematic reviews. Furthermore, when conducting reproducibility analysis of reported quantitative results of primary studies, researchers often rely on the original data provided by the original authors (e.g., Artner et al., 2020; Hardwicke et al., 2018, 2021). In the context of research synthesis, this would involve an assessment of the reproducibility of the (quantitative) synthesis of included studies stage using the primary data already coded by the original authors. This puts the focus of the assessment at factors such as data availability, reusability of the data, challenges in reconstructing the original analysis scheme, and potential reporting errors.

1.4 Aims of this dissertation

As discussed in previous sections, research synthesis has transitioned from a narrative and subjective approach to a rigorous scientific enterprise. As any scientific output, research synthesis results and conclusions are expected to provide sound and rigorous evidence. This is particularly notable in the case of evidence synthesis, which is often placed at the top of hierarchy of evidence, exerting significant influence on policymaking, social practices, and healthcare decisions. Furthermore, as discussed above, the recent credibility crisis within psychological research has highlighted the value of monitoring research practices in order to identify issues and drive improvements. However, as previously stated, research synthesis has received relatively less attention in this regard. Therefore, the current dissertation aims to explore possible issues, challenges, or poor practices in the context of reproducibility of research synthesis projects.

Specifically, the current dissertation empirically assesses several aspects through the following three chapters. Firstly, it examines the transparency and reproducibility-related reporting practices of research syntheses. Secondly, it investigates the reproducibility of reported results in quantitative synthesis (meta-analysis). Lastly, it evaluates the statistical power of random-effects meta-analyses. These assessments are conducted on a random sample of published research syntheses focusing on clinical psychological interventions.

Next, a brief summary of each of the included studies is provided.

1.4.1 A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020).

In this study, we attempted to empirically assess the prevalence of transparency and reproducibility-related reporting practices in published meta-analyses from clinical psychology by examining a random sample of 100 meta-analyses. Our purpose was to identify the key points that could be improved, with the aim of providing some recommendations for carrying out reproducible meta-analyses. We conducted a meta-review of meta-analyses of psychological interventions published between 2000 and 2020. We searched PubMed, PsycInfo and Web of Science databases. A structured coding form to assess transparency indicators was created based on previous studies and existing meta-analysis guidelines. We found major issues concerning: completely reproducible search procedures report, specification of the exact method to compute effect sizes, choice of weighting factors and estimators, lack of availability of the raw statistics used to compute the effect size and of interoperability of available data, and practically total absence of analysis script code sharing. Based on our findings, we conclude with recommendations intended to improve the transparency, openness, and reproducibility-related reporting practices of meta-analyses in clinical psychology and related areas.

1.4.2 Reproducibility of published meta-analyses on clinical psychological interventions.

In this study, 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.

1.4.3 Statistical power of random-effects meta-analyses on clinical psychological interventions.

Underpowered studies are ubiquitous in psychology and related disciplines. Meta-analysis can help alleviate this problem, increasing the statistical power by combining the results of a set of primary studies. However, this is not necessarily true when we use a random-effects model, which is currently the predominant approach when carrying out meta-analyses. In this study, we examined the statistical power of a sample of 141 random-effects meta-analyses on the effectiveness of clinical psychological interventions.

Additionally, we compared the estimated power of these meta-analyses with the power of the individual studies that comprised them. To do so, we used different analytical approaches and a Monte Carlo approach. The statistical power of random-effects meta-analyses was computed under different scenarios of true effect size and level of heterogeneity. Our results show that under certain scenarios, random-effects meta-analytic statistical testing is underpowered, even showing a lower statistical power than the average or maximum statistical power of included primary studies. Overall, these scenarios were characterised by high heterogeneity and a low number of included studies. While this pattern is expected, our findings show the steepness of this drop in statistical power. These results are discussed in light of statistical and conceptual basis of random-effects meta-analysis.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., & Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions*, *25*(1), 12–37. https://doi.org/10.1177/0163278702025001003

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*(2), 234–237. https://doi.org/10.1037/0021-9010.61.2.234

Cochran, W. G. (1937). Problems Arising in the Analysis of a Series of Similar Experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 102–118. https://doi.org/10.2307/2984123

Cochran, William G. (1954). The Combination of Estimates from Different Experiments. *Biometrics*, *10*(1), 101–129. https://doi.org/10.2307/3001666

Cochran, William G., & Carroll, S. P. (1953). A Sampling Investigation of the Efficiency of Weighting Inversely as the Estimated Variance. *Biometrics*, *9*(4), 447–459. https://doi.org/10.2307/3001436

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. https://doi.org/10.1037/h0045186

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*(2), 291–302. https://doi.org/10.2307/1170314

Cooper, H. M. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (Fifth Edition). SAGE.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis* (3rd edition). Russell Sage Foundation.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

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

Fisher, R. A. (1925). *Statistical Methods for Research Workers* (1sr edition). Oliver and Boyd.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th edition). Oliver and Boyd.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate ψ. *Journal of Personality and Social Psychology*, *103*(6), 933–948. https://doi.org/10.1037/a0029709

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

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

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

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., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359–369. https://doi.org/10.1037/0033-2909.88.2.359

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

Higgins, J. P. T., & Collaboration, C. (Eds.). (2020). *Cochrane handbook for systematic reviews of interventions* (Second edition). Wiley-Blackwell.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Hunt, M. (1997). *How Science Takes Stock: The Story of Meta-Analysis*. Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610442961

Jackson, G. B. (1980). Methods for Integrative Reviews. *Review of Educational Research*, *50*(3), 438–460. https://doi.org/10.3102/00346543050003438

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kitzes, J. (2018). The Basic Reproducible Workflow Template. In J. Kitzes, D. Turek, & F. Deniz (Eds.), *The Practice of Reproducible Research*. University of California Press.

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

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

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

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Maassen, E., van Assen, M. A. L. M., 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

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, *11*(6), 725–742. https://doi.org/10.1002/jrsm.1452

Mayo-Wilson, E., & Grant, S. P. (2019). Transparent Reporting: Registrations, Protocols, and Final Reports. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analysis* (3rd edition). Russell Sage Foundation.

McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thompson, H. J., Johnston, R. V., & Thomas, J. (n.d.). Defining the criteria for including studies and how they will be grouped for the synthesis. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (version 6.3 (updated February 2022)). Cochrane, 2022.

McKenzie, J. E., & E., B., Sue. (n.d.). Synthesizing and presenting findings using other methods. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (version 6.3 (updated February 2022)). Cochrane, 2022.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting Transparency in Social Science Research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

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

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., McKenzie, J. E., 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., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pearson, Karl. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Pearson, Karl. (1904). Report on Certain Enteric Fever Inoculation Statistics. *British Medical Journal*, *2*(2288), 1243–1246. https://doi.org/10.1136/bmj.2.2288.1243

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, *25*, 379–410. https://doi.org/10.1093/biomet/25.3-4.379

Pillemer, D. B. (1984). Conceptual Issues in Research Synthesis. *The Journal of Special Education*, *18*(1), 27–40. https://doi.org/10.1177/002246698401800105

Rice, K., Higgins, J. P. T., & Lumley, T. (2018). A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society Series A*, *181*(1), 205–227.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, *12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*.

Schmid, C. H., Stijnen, T., & White, I. R. (Eds.). (2020). *Handbook of meta-analysis* (First edition). Taylor and Francis.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2023). *[109] Data Falsificada (Part 1): "Clusterfake"*.

Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significanceor Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams Jr., R. M. (1949). *The American soldier: Adjustment during army life.* (pp. xii, 599). Princeton Univ. Press.

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

Valentine, J. C. (2019). Incorporating Judgments About Study Quality into Research Syntheses. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analysis* (3rd edition). Russell Sage Foundation.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

White, H. D. (2019). Scientific communication and literature retrieval. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analysis* (3rd edition). Russell Sage Foundation.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*(7375), 7–7. https://doi.org/10.1038/480007a

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, *48*(2), 156–158. https://doi.org/10.1037/h0059111

Chapter 2

A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020)[1]

Abstract

Meta-analysis is a powerful and important tool to synthesize the literature about a research topic. Like other kinds of research, meta-analyses must be reproducible to be compliant with the principles of the scientific method. Furthermore, reproducible meta-analyses can be easily updated with new data and reanalysed applying new and more refined analysis techniques. We attempted to empirically assess the prevalence of transparency and reproducibility-related reporting practices in published meta-analyses from clinical psychology by examining a random sample of 100 meta-analyses. Our purpose was to identify the key points that could be improved with the aim to provide some recommendations to carry out reproducible meta-analyses. We conducted a meta-review of meta-analyses of psychological interventions published between 2000 and 2020. We searched PubMed, PsycInfo and Web of Science databases. A structured coding form to assess transparency indicators was created based on previous studies and existing meta-analysis guidelines. We found major issues concerning completely reproducible search procedures report, specification of the exact method to compute effect sizes, choice of weighting factors and estimators, lack of availability of the raw statistics used to compute the effect size and of interoperability of available data, and practically total absence of analysis script code sharing. Based on our findings, we conclude with recommendations intended to improve the transparency, openness and reproducibility-related reporting practices of meta-analyses in clinical psychology and related areas.

[1] **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.

2.1 Introduction

Systematic reviews and meta-analyses are commonly ranked among the most relevant sources of scientific evidence on the effectiveness of healthcare interventions (Evans, 2003), and therefore provide a powerful tool to evidence-based healthcare practice. Importantly, the validity of the conclusions drawn from a meta-analysis depends on the methodological quality and rigor of the primary studies (Nuijten et al., 2015; van Assen et al., 2015).

The last decade has revealed significant problems in terms of replicability and reproducibility in psychological research, leading to the so-called 'replication crisis' (McNutt, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). In this paper, by replicability, we mean that a previous conclusion will be supported by novel studies that address the same question with new data, and, by reproducibility, we refer to obtaining the exact same previous result applying the same statistical analysis to the same data (Asendorpf et al., 2013; Epskamp, 2019).

Several efforts have been made to evaluate the replicability of findings from psychology and related fields (e.g., Hagger et al., 2016; Klein, 2014; Open Science Collaboration, 2015). A number of methodological issues, questionable research practices, and reporting biases have been suggested as potential explanations for failed replication attempts (Ioannidis, 2005; Johnson et al., 2017; Schmidt & Oh, 2016; Simmons et al., 2011; Stanley et al., 2018; Szucs & Ioannidis, 2017). In this context, meta-research has emerged as an approach 'to investigate quality, bias, and efficiency as research unfolds in a complex and evolving scientific ecosystem' (Hardwicke et al., 2020a, p. 12; Ioannidis, 2018). This 'research on research' aims to help identify the key points that could be improved in research and reporting practices.

Different concerns about the reproducibility of published meta-analyses have also emerged. Gøtzsche et al. (2007) re-computed the primary effect sizes from 27 meta-analyses, finding problems in 10 of them. Tendal et al. (2009) re-computed the primary effect sizes and summary meta-analytic estimates re-extracting the relevant primary statistics by independent coders, finding substantial inconsistencies. In a similar way, Tendal et al. (2011) found that multiplicity of effect sizes in primary studies can lead to different meta-analytic conclusions depending on how such multiplicity is addressed. Lakens et al. (2017) struggled to reproduce a set of meta-analyses due to lack of access to raw data and incomplete reporting of the methodology followed. Kvarven et al. (2020) compared the results of published meta-analyses to large-scale replications on the same topic, finding significant differences in effect sizes for 12 out of the 15 pairs. And last, Maassen et al., (2020) found a number of challenges to reproduce the calculation of effect sizes based on the information reported by the original authors of each meta-analysis.

Of note, carrying out a meta-analysis involves a multi-decision process from the literature search to the statistical analysis, and only if such decisions are clearly stated will the meta-analysis be reproducible by an independent research team. Open science initiatives are a major point here: pre-registration, sharing open material and data, and sharing open analysis scripts offer several benefits (Federer et al., 2018; Hardwicke & Ioannidis, 2018b; Nosek & Lindsay, 2018; Nelson et al., 2018; Nosek et al., 2015; Nosek et al., 2019; Popkin, 2019). The importance of promoting and adopting open science practices in meta-analysis has been increasingly recognised in recent years (Lakens et al., 2016; Moreau & Gamble, 2020; Pigott & Polanin, 2020). For instance, pre-registered meta-analyses avoid to some extent practices such as selective inclusion or reporting of results (Page et al., 2013). Additionally, open meta-analytic data sharing offers several benefits related to efficiency in scientific development and reproducibility or robustness checking. Full, machine-readable availability of meta-analytic data allows for: easy updating, reusability for new purposes, reanalysis with different or novel analysis techniques and quick checking of possible errors. Along with the availability of meta-analytic data, open script code sharing allows for easy analytical reproducibility checking and involves a straightforward statement of the analytic methods applied. All these points are particularly relevant in the context of meta-analysis given that meta-analysis claims may have a strong impact on policy making or healthcare practices. In addition, meta-analyses should keep the results updated as new primary evidence emerges. It is important to note that there is no a single perspective concerning which analytic methods should be applied in meta-analysis, so that novel analytic methods are regularly being developed. Applying such novel techniques to published data could be enlightening.

The last years have seen a proliferation of reviews assessing the prevalence of transparency and reproducibility-related practices in primary studies. A common finding across such reviews is the lack of transparency in the reporting of key indicators for reproducibility. Some of these reviews examined broad research disciplines such as biomedical sciences (Iqbal et al., 2016; Wallach et al., 2018), social sciences (Hardwicke et al., 2020c), and psychology (Hardwicke et al., 2020b). In the meta-analytic arena, Polanin et al. (2020) assessed the compliance with transparency and reproducibility-related practices of all meta-analyses published in Psychological Bulletin, finding poor adherence to these guidelines. This restriction to a specific journal arguably yielded a pool of high-quality meta-analyses, but it remains unclear whether the patterns observed can be generalized to other journals with different editorial guidelines and requirements. While Polanin et al.'s (2020) approach provides an overview of the reporting quality of meta-analyses across a wide range of scientific topics, it also makes it difficult to characterize the reporting pattern in a specific research area.

2.1.1 Purpose

In this study we empirically assessed the prevalence of transparency and reproducibility-related practices in published meta-analyses on clinical psychological interventions examining a random sample of 100 meta-analyses. Our purpose was to identify the key points that could be improved in the field of clinical psychology and to produce some recommendations accordingly. We selected the area of effectiveness of clinical psychological interventions for three main reasons. First, we intended to offer recommendations focused on a specific research topic, since transparency and openness practices might vary across research areas. Second, meta-analysis on the effectiveness of clinical psychological interventions is one of the types of meta-analysis most frequently published in psychological research. Third, meta-analyses on the effectiveness of clinical psychological interventions have an important impact on clinical practice and policy making.

2.2 Method

2.2.1 Design

This is a meta-review, that is, a kind of umbrella review that can be defined as a methodological systematic review of meta-analyses (Biondi-Zoccai, 2016).

2.2.2 Identification and selection of studies

Published meta-analyses of clinical psychological interventions were identified conducting a systematic electronic search in PubMed, Scopus, and the core collection of Web of Science. The search was carried out on the 22nd of January 2020. The full search strategies followed in each database are available in Appendix 2A. Articles were included if the following criteria were met: (a) At least one meta-analysis focused on the effectiveness of psychological intervention/s was reported; (b) publication year after 1999; (c) the effect size index was a mean difference or a standardized mean difference, and (d) written in English or Spanish. Individual participant data meta-analyses and network meta-analyses were excluded from this study.

All records identified by the electronic search were downloaded in bibliographic format and duplicates were removed using the R package '*revtools*' (Westgate, 2019), first by exact match from DOIs, and subsequently by fuzzy matching from titles. All bibliographic files (the outputs of electronic search and the output of unique references) and the script code used to remove duplicates are available at: https://osf.io/xg97b/. Unique references were uploaded to the open-source program '*abstrackr*' (Wallace et al., 2012) for the screening. The titles and abstracts of the unique references were assessed by one author (RLN), and references that were clearly ineligible were excluded at this stage. When the information presented in title and abstract was insufficient, the full-text records were evaluated independently by two authors (RLN and MRA), with a third author (JSM or JLL) getting involved to resolve any disagreements. Appendix 2A presents a flow chart summarising the screening process.

2.2.3 Sampling

A total of 664 meta-analyses were identified by the electronic search and screening process. Of these, 100 were randomly selected using a random number generator between 1 and the total number of meta-analyses included, setting up a certain seed to guarantee the reproducibility of the process. Appendix 2A presents two overlapping histograms displaying the distribution of the year of publication for the included meta-analyses and for the selected random sample. In order to compare the two observed distributions, the Kolmogorov-Smirnov test was performed. Equivalence was found between both distributions (D = .104, p = .299).

2.2.4 Procedure and data extraction

A structured coding form was created based on previous studies (Hardwicke et al., 2020c; Iqbal et al., 2016; Koffel & Rethlefsen, 2016; Wallach et al., 2018) and existing meta-analyses guidelines (Liberati et al., 2009; Pigott & Polanin, 2020). The coding form is available at: https://osf.io/2dzmk/.

Items were grouped into nine different categories: (a) study ID and study characteristics (items 1-7); (b) pre-registration, protocol, and the statement of compliance to guidelines (items 7-13); (c) identification and selection of studies (items 14-23); (d) data collection process (items 24-29); (e) effect or summary measures (items 30-35); (f) statistical methods (items 36-46); (g) data and script analysis availability (items 47-59); (h) conflict of interest and funding statement (items 60-61), and (i) access format of the paper (item 62).

At a first stage, the coding form items were tested in a pilot coding. Four authors (RLN, MRA, JSM and JLL) independently applied the coding form to a random sample of 5 meta-analyses. Subsequently, in a series of meetings, disagreements between the coders were resolved by discussion until consensus was reached. During this process, items were modified or refined where necessary.

Next, two authors (RLN and MRA) independently applied the coding form to the 100 meta-analyses randomly selected. The coding form was applied between April 3rd, 2020 and May 29th, 2020. Discrepancies between the two coders were resolved by discussion and review of the relevant materials. The three datasets (coder 1, coder 2 and consensus data) are available at: https://osf.io/xg97b/. Inter-coder agreement was assessed with Cohen's Kappa coefficient, for close-ended items, using the R package '*irr*' (Gamer et al., 2019). The resulting values ranged between .55 and 1, with only two items yielding values below .6 (item 16 and 55, see Appendix 2B).

In addition, the format used to share each kind of raw data available was coded a posteriori[2], given the implications of this aspect for the efficient reusability of the data. Thus, six subitems paired with items 50-55 were added. The formats were categorised as interoperable or not (Bek, 2019; Wilkinson, 2016) based on two criteria: format that allows to easily manipulate and read the values for open-source statistical software and proprietary/non-proprietary format.

2.2.5 Analysis

First, we examined how often each of the indicators was reported across meta-analyses. For each proportion, we calculated 95% confidence intervals based on the Wilson score interval (Wilson, 1927) for binomial items and on the Sison-Glaz method (Sison & Glaz, 1995) for multinomial items, using the R package '*DescTools*' (Signorell et al., 2020).

Furthermore, we explored possible associations using binary logistic regression, with publication year (item 4), pre-registration (item 7) and use of reporting guidelines (item 12) as predictors, and the following dichotomous (or dichotomized by removing the "Other" category) indicators as dependent variables: items 15 to 20, 22 to 32, 34, 36, 38 to 42, 44, 50 to 55. We started fitting single predictor models to observe unadjusted associations, and then switched to multiple regression models introducing all three predictors to explore the associations for each predictor controlling for the others. We quantified the strength of the associations by calculating odds ratios and 95% confidence intervals based on profile likelihood. Despite the large number of contrasts performed, we did not introduce any corrections for multiple comparisons due to the exploratory nature of our analyses.

Preparation of data and all figures presented in this paper was accomplished using the collection of R packages '*tidyverse*' (Wickham et al., 2019). All the script codes used to handle and analyse the data are openly available at: https://osf.io/xg97b/.

[2] Only cases that provided data from different sources than the article itself (previously coded in item 48) were re-reviewed. For cases that only provided data in the article itself (item 48 = 2), "pdf" was imputed for each type of data previously coded as available (see script analysis code available at: https://osf.io/a7zth/).

2.3 Results

The total of 664 included meta-analyses were published between 2000 and 2020 (median = 2015), whereas publication year for the selected random sample of 100 meta-analyses ranged between 2001 and 2020 (median = 2016).

2.3.1 Pre-registration, guidelines and conflict of interest

Of the 100 meta-analyses examined, 19 (see Figure 1a) stated that there was a pre-registration of the study; of these, 13 (68%, Figure 1b) allocated their pre-registration in PROSPERO, 3 (16%) in Cochrane library, 1 (5%) in OSF, 1 (5%) in UMIN-CTR, and 1 (5%) internally at a national agency. Conversely, 78 out of the 100 meta-analyses in our random sample did not include any statement on pre-registration, whereas 2 stated that there was no pre-registration and 1 mentioned pre-registration of a different project. Only 17 out of the 100 meta-analyses included a link or a unique ID to locate an accessible protocol (Figure 1c).

With regards to the statement of compliance to guidelines (Figure 1d), 70 out of the 100 meta-analyses did not mention following any reporting guideline, whereas 27 stated to follow PRISMA and 3 stated to follow other guidelines (MARS in 2 studies and QUOROM in 1).

Funding sources and competing interests could be a potential source of bias. Of the 100 meta-analyses reviewed, 13 (see Figure 1e) stated one or more conflicting interests, 51 stated that there were no conflicting interests, and 36 did not include a conflict-of-interest statement. With regards to funding, 38 meta-analyses (see Figure 1f) failed to include a funding statement, whereas 38 declared public funding sources, 3 mentioned private sources, 1 declared both public and private sources, and 20 stated that no funding was provided. About accessibility, 29 of the 100 meta-analyses had no publicly available version; of these, 13 stated that public funding was provided.

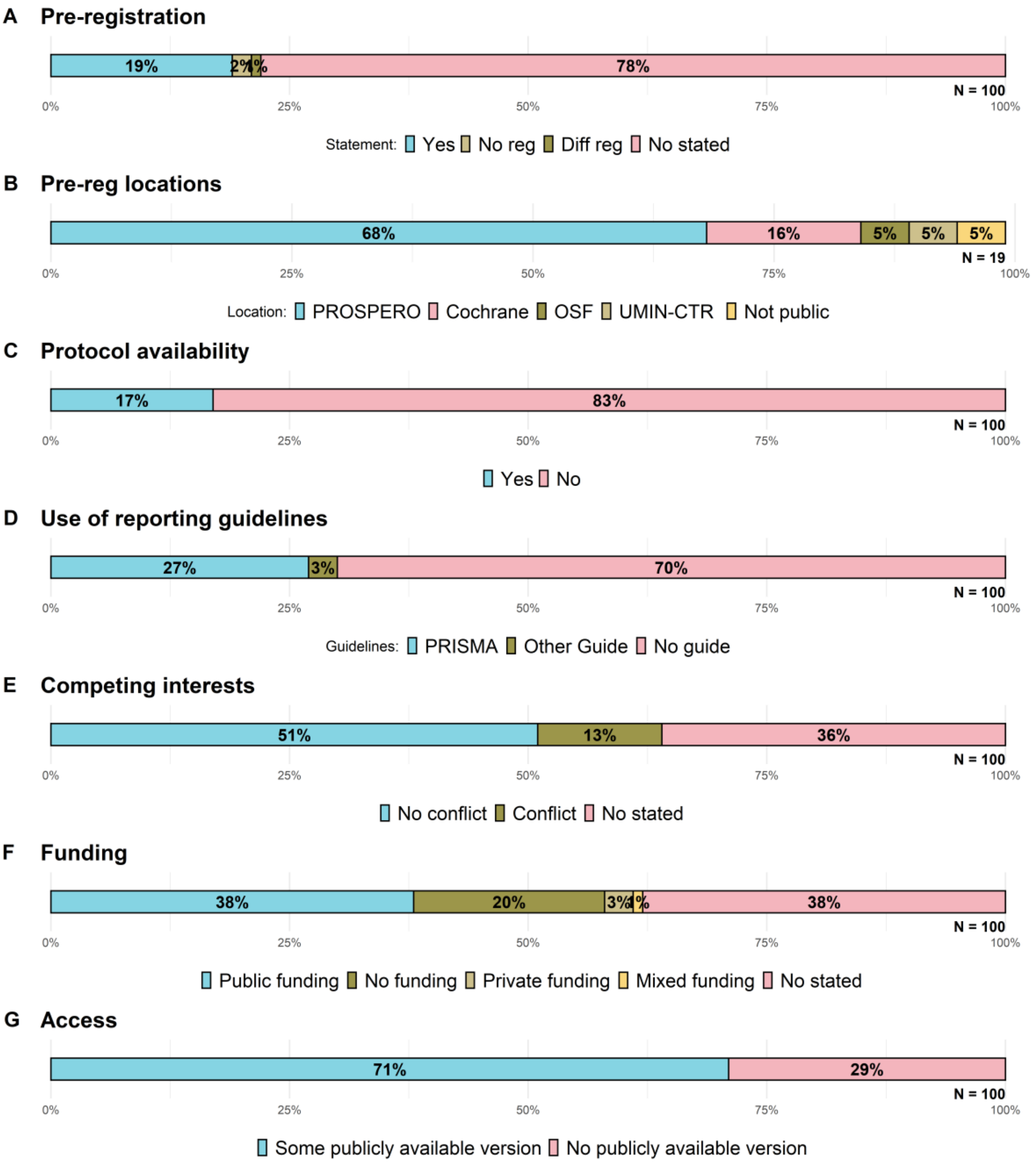


Figure 1. Percentage of (a) meta-analysis pre-registered, (b) pre-registration locations, (c) protocol availibility, (d) guidelines adherence, (e) competing interest statements, (f) funding statements, and (g) accesibility of meta-analyses. N indicates total number of meta-analyses assessed for each indicator.

2.3.2 Systematic review methods

Eligibility criteria and literature search

Detailed and complete reporting of the search and screening procedures allows the assessment of the quality of the procedure and facilitates replication. We excluded 1 meta-analysis because it consisted of a re-analysis of a previous meta-analysis. Thus, this meta-analysis was excluded from the analysis of the items concerning electronic search (item 14 to 20). All the remaining 99 meta-analyses specified the electronic databases consulted (Figure 2a); of these, 66 (67%) specified the year for first date searched (including database inception); 69 (70%) indicated the electronic search limits used; 84 (85%) specified the month and year of the electronic search; 93 (94%) included the search terms used; 63 (64%) reported the full search strategy (exact terms and the Boolean connectors). However, only 37 reported all these details combined, which is required for the electronic search to be completely reproducible; 86 (87%) declared having used additional search methods as follows: 78 (91%) used additional backward searches of reference lists of identified articles or relevant previous reviews, 29 (34%) used additional hand searches of relevant websites, conferences papers, relevant journals, etc., 23 (27%) contacted experts, 9 (10%) consulted Google Scholar, and 5 (6%) used additional forward searches by citation tracking.

Among the 100 meta-analyses examined, 96 (Figure 2a) specified the eligibility criteria and 82 described the screening process.

Data collection process

The data collection process should be detailed, including the methods for dealing with missing data and for assessing risk of bias in the included studies, so that the accuracy of the extracted data and its validity can be evaluated. Of the 100 meta-analyses, 68 (see Figure 2b) described details about the collection process of study characteristics; out of these, 61 (90%) conducted double coding, of which 21 (34%) reported inter-coder agreement values. Also, 77 out of the 100 meta-analyses listed all variables for which data were sought, 42 described at least one method to deal with missing data (such as statistical imputation, request to authors, etc.), and 77 described methods to assess risk of bias in included studies.

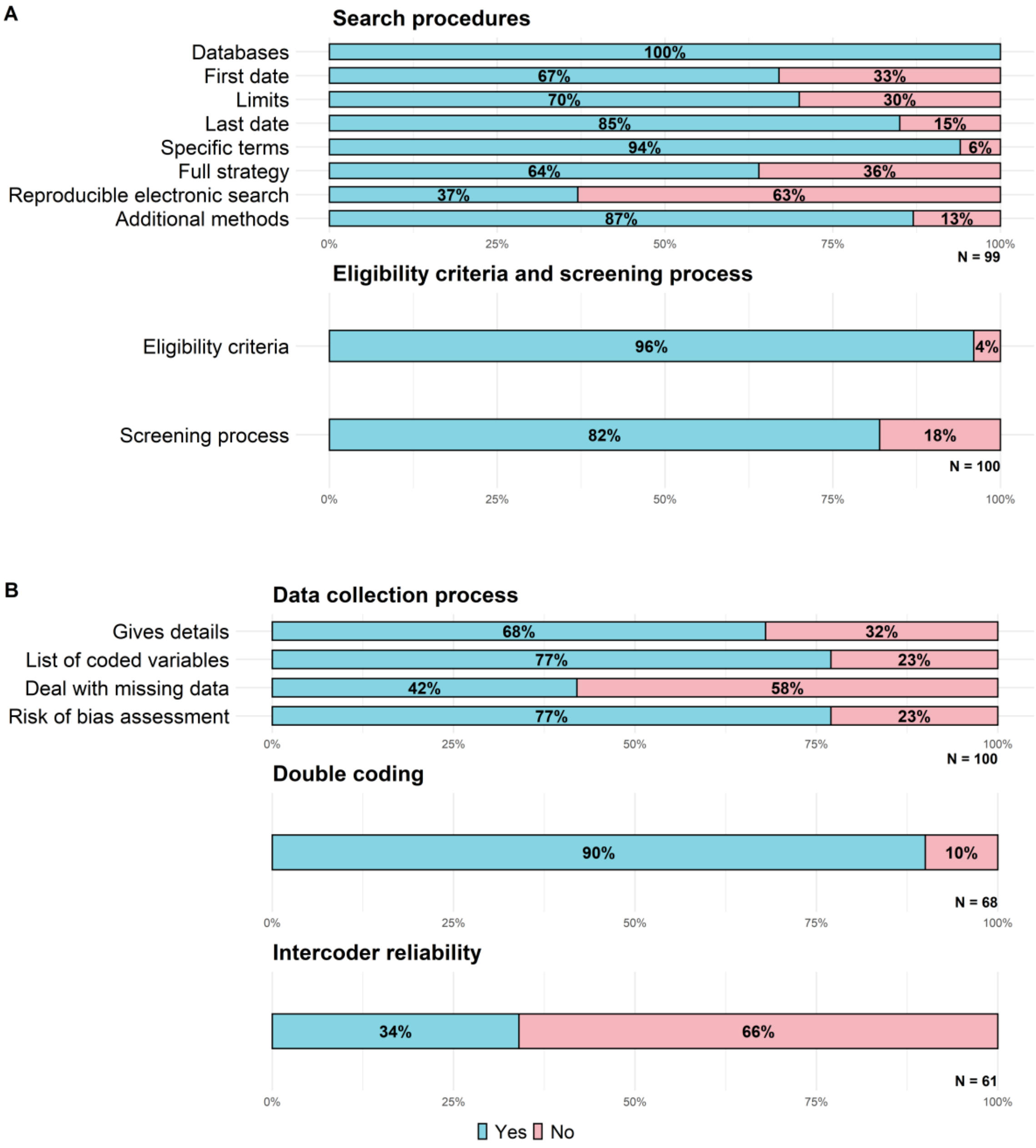


Figure 2. Percentage reported of systematic review methods by (a) eligibility criteria and literature search, and (b) data collecion process, showing different indicators for each category. N indicates total number of meta-analyses assessed for each indicator.

2.3.3 Meta-analysis methods

Effect measures

Identifying the effect measure used and specifying the method to calculate it is crucial due to the existence of many different effect size measures as well as several approaches to calculate some of them (Hoyt & Del Re, 2018; Rubio-Aparicio et al., 2018). The majority of the 100 meta-analyses reported the effect measure used in the synthesis (93% see Figure 3a), however, the majority of these did not specify in detail which formula was used to compute it (85%).

Multiplicity of results in trial reports leads to statistical dependency if the multiple effect estimates from the same study are based (at least partially) on the same participants, and ignoring it may result in underestimation of standard errors and erroneous statistical conclusions (Bender et al., 2008; López-López et al., 2018; Tendal et al., 2011). About half of the meta-analyses (54%) described at least one method to deal with multiplicity, including random selection, averaging, decision rules or using advanced meta-analytic methods to model or account for it (López-López et al., 2018). About a third (33%) of the meta-analyses described sensitivity analyses to assess the effect of outliers.

Synthesis and analysis methods

The choice of statistical model and meta-analytic method may have an impact on the results and conclusions, hence the importance of reporting a detailed description of the statistical analysis approach (Langan et al., 2015; Sánchez-Meca et al., 2013; Schmidt et al., 2009). The vast majority of the 100 meta-analyses stated the statistical model assumed for the synthesis process (92%, see Figure 3b), with most of them assuming a random-effects model (87, 95%); however, very few of those meta-analyses stated the estimation method of the heterogeneity variance, τ^2 (11, 13%). Furthermore, of the total of 100 meta-analyses, only 30 stated the weighting factor used, whereas 85 mentioned methods to assess heterogeneity. Moreover, 65 meta-analyses described methods to assess the influence of possible moderator variables, but only 22 of these (34%) specified the statistical model assumed for the moderator analyses.

Additionally, 73 out of the 100 meta-analyses stated having used at least one method to assess reporting biases (including publication bias); of these, 61 (84%) reported a funnel plot, 34 (47%) applied the trim-and-fill method, 31 (42%) used the Egger test, 24 (33%) applied some form of the fail-safe-N method, 13 (18%) used the Begg and Mazumdar test, and only one used PET-PEESE and p-uniform methods.

Most meta-analyses identified the software used to carry out the statistical analyses (89%); of these, 38 (43%) used Comprehensive Meta-Analysis, 24 (27%) used Review Manager, 20 (22%) used STATA, 12 (13%) used R, 8 (9%) used SPSS, and 6 (7%) used other software.

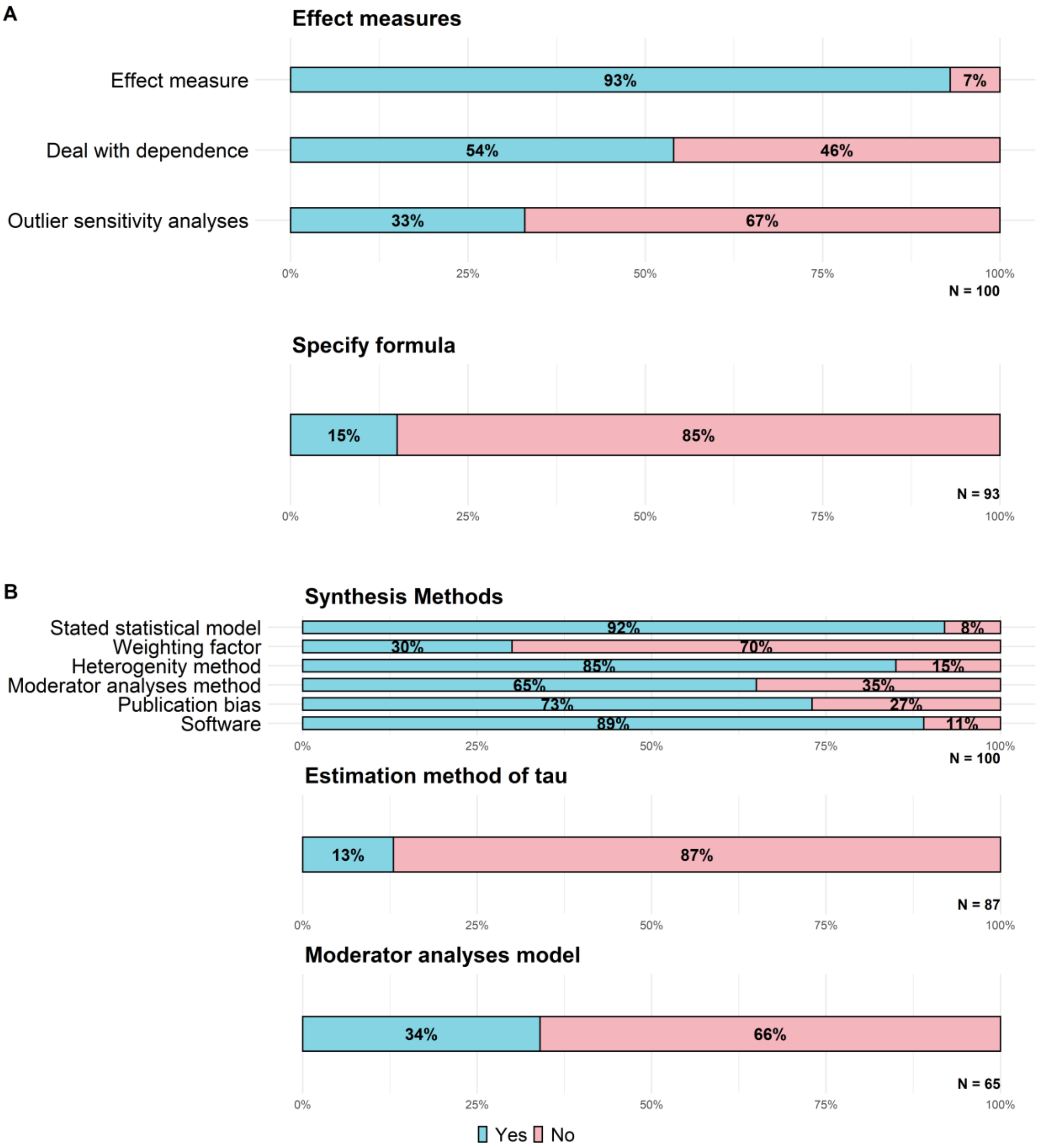


Figure 3. Percentage reported of meta-analysis methods by (a) effect measures, and (b) synthesis and analysis methods, showing different indicators for each category. N indicates total number of meta-analyses assessed for each indicator.

2.3.4 Data and analysis script availability

The unit of analysis of a meta-analysis is usually the primary study, so when we talk about data availability, we typically refer to the summary level data (e.g., effect sizes) from each primary study included in each meta-analysis. In systematic reviews and meta-analyses, it is common to report the characteristics of the included studies, as well as, through table or forest plots, the individual effects measures. The vast majority of the meta-analyses we examined (98%, see Figure 4a) reported at least some raw data; of these, 93 reported some raw data in the paper itself. Furthermore, 31 meta-analyses included raw data in supplementary files or appendices, 4 stated that there was some raw data upon request, 1 shared data using an institutional webpage, and 1 using https://osf.io/.

Of the 98 meta-analyses for which some raw data was available, All the meta-analyses (see Figure 4b) identified the primary study associated with the data, only 3 in interoperable format; 89 reported the primary study comparator (e.g., treatment-as-usual, waitlist, other intervention…), only 3 in interoperable format; 82 reported the primary effect sizes combined, only 3 in interoperable format: 69 reported the sample sizes of the groups compared in the primary studies, only 3 in interoperable format: 29 reported the statistics used to compute primary effect sizes, only 2 in interoperable format: and 70 reported the coded moderator variables, only 3 in interoperable format

Data script availability refers to detailed step-by-step descriptions of the analyses carried out (e.g., SPSS syntax, R code etc.). Availability of the analysis code, along with the data shared, enables to check computational reproducibility of the reported results. Unfortunately, only 1 of the meta-analyses we examined (see Figure 4c) mentioned that the analysis script code was available (through an OSF link).

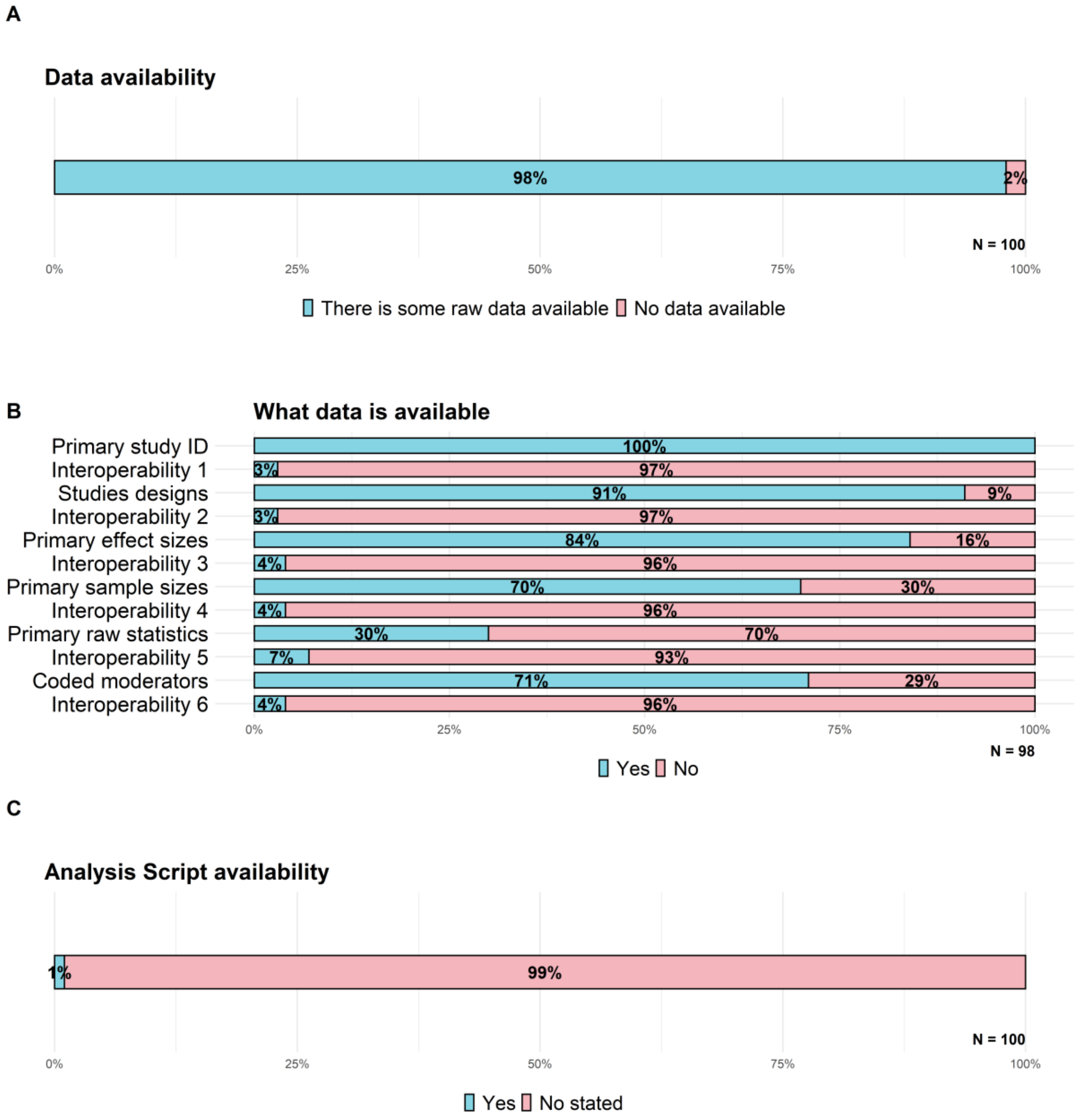


Figure 4. Percentage of (a) meta-analysis that reported some raw data, (c) meta-analysis that shared the analysis script code, and (b) what data were available and if these were in interoperable formats; each interoperability bar corresponds to the primary data represented over it. N indicates total number of meta-analyses assessed for each indicator.

2.3.5 Associations between year, pre-registration or guidelines adherence and transparency and reproducibility-related reporting

Several logistic regression models were fitted, for space-saving reasons only a selection of the results is presented in this section. The full results are available at: https://osf.io/9xsg2/

Table 1 presents the odds ratio and 95% CI of the main results of simple and multiple models. Taking into consideration the results of the simple and multiple models, publication year was a significant predictor of the inclusion of a description of the screening process (*OR* = 1.29 [95%CI: 1.12-1.54], the statistical model assumed (*OR* = 1.29 [95%CI: 1.08-1.60]), the methods to assess reporting biases (*OR* = *1*.19 [95%CI: 1.06-1.35]), and the software used (*OR* = 1.19 [95%CI: 1.04-1.39]), with more recent studies providing a more detailed description of the methods used. Moreover, pre-registered meta-analyses were more likely to specify the year for first date searched (*OR* = 13.27 [95%CI: 2.32-253.59]) and following reporting guidelines such as PRISMA was associated with a more complete report of the full search strategy (*OR* = 3.20 [95%CI: 1.07-11.08]) and the methods used for assessing risk of bias of the individual studies (*OR* = 6.50 [95%CI: 1.12-124.12]).

Table 1.

Odds ratio and 95% CI between predictors and transparency and reproducibility-related indicators

Indicator	Year		Pre-registration		Guideline adherence statement	
	Simple	Multiple	Simple	Multiple	Simple	Multiple
Specify the year for first date searched	1.06 [0.96-1.17]	1.04 [0.93-1.17]	**12.00 [2.30-221.22]**	**13.27 [2.32-253.59]**	1.24 [0.50-3.25]	.64 [0.21-1.93]
Report the full search strategy	1.1 [1.00-1.22]	1.05 [0.94-1.17]	2.50 [0.82-9.38]	1.41 [0.40-5.76]	**4.08 [1.49-13.20]**	**3.20 [1.07-11.08]**
Specify the eligibility criteria operatively	**1.23 [1.01-1.52]**	1.19 [0.95-1.50]				
Describe the screening process	**1.32 [1.16-1.53]**	**1.29 [1.12-1.54]**			**9.30 [1.77-171-82]**	2.44 [0.37-48.35]
List all variables for which data were sought	**1.15 [1.03-1.29]**	1.12 [0.99-1.26]	2.97 [0.90-13.55]	1.56 [0.33-11.25]	**3.60 [1.11-16.25]**	2.27 [0.61-11.18]
Describe methods used for assessing risk of bias of individual studies	**1.17 [1.05-1.32]**	1.10 [0.97-1.25]			**13.29 [2.57-244.28]**	**6.50 [1.12-124.12]**
Identify the statistical model assumed	**1.23 [1.06-1.45]**	**1.29 [1.08-1.60]**			1.31 [0.28-9.34]	0.25 [0.03-2.29]

Identify the estimation method of τ^2	1.15 [0.96-1.47]	1.06 [0.89-1.34]	**4.55 [1.16-17.42]**	3.12 [0.71-13.41]	3.18 [0.88-12.03]	1.97 [0.47-8.43]
Describe any methods to assess reporting biases (including publication bias)	**1.16 [1.04-1.29]**	**1.19 [1.06-1.35]**	3.79 [0.99-25.08]	4.73 [0.97-38.21]	0.81 [0.32-2.14]	**0.29 [0.09-0.94]**
Mention the software used to carry out the statistical analyses	**1.20 [1.05-1.38]**	**1.19 [1.04-1.39]**	2.54 [0.44-48.04]	1.58 [0.19-35.96]	2.07 [0.49-14.13]	0.99 [0.17-8.25]
Statistics used to compute the effect are size available	1.09 [0.98-1.24]	1.05 [0.93-1.2]	2.08 [0.72-5.86]	1.38 [0.43-4.26]	**2.58 [1.03-6.48]**	2.04 [0.74-5.66]

Note: Odds ratio and CIs non-interpretable due to separation were omitted; Odds ratio 95% CI is presented in brackets. Bolded values indicated CIs that do not contain the null value.

2.3.6 Key points

The key points identified where a substantial lack of transparency was found concerning the potential reproducibility of the meta-analyses examined are summarized in Table 2. Other aspects related to the promoting transparency (i.e., well-stablished reporting guidelines adherence) and to the prevention of result-based bias (i.e., pre-registration) are summarized in Table 3.

2.4 Discussion

The main aim of this study was to analyse the prevalence of transparency and reproducibility-related practices in meta-analyses on the effectiveness of clinical psychological interventions. A random sample of published meta-analyses on the effectiveness of clinical psychological interventions was reviewed. Additionally, the relationship between publication year, pre-registration, and guidelines adherence and different indicators was assessed. A lack of transparency in key aspects for the reproducibility of meta-analyses was found.

Regarding pre-registration, the 19% of pre-registered meta-analyses found in our meta-review is substantially higher than findings from previous studies mainly focused on primary research (Hardwicke et al., 2020b, 3%; Hardwicke et al., 2020c, 0%;) and higher than that found in a previous study focused on meta-analyses (Polanin et al., 2019, 2%). However, the existence of a pre-registration was not shown to be associated with an increased reporting of information related to the potential reproducibility of the meta-analysis, except for the specification of the year for first date searched and, to a minor extent, for identification of the estimation method of the heterogeneity variance. The majority of identified pre-registrations were allocated in specialized repositories such as PROSPERO, and these were submitted through a structured form. Hence, relevant information, identified in this study as poorly reported, could be explicitly requested, such as: full search strategy, estimation method of the heterogeneity variance, or the formula used to compute the effect measure. As pointed in Table 3, it is worth noting that pre-registration is compatible with flexibility, allowing flexibility tracking. Regarding guidelines adherence statements, only 30 of the 100 meta-analyses stated the use of reporting guidelines. Adherence statements to guidelines was associated to higher reporting of the full search strategy, full description of the methods used for assessing risk of bias of individual studies and, to a minor extent, better description of the screening process, coded variables, and the statistics used to compute the effect measure. The suboptimal adherence to many items of PRISMA guidelines have been studied in previous studies (Page et al., 2017). An update of PRISMA has recently been published (Page et al., 2021), including new recommendations and changes relevant to some of the aspects examined in this study.

The reporting of search strategy elements in clinical psychology was found to be better than in other areas (Koffel & Rethlefsen, 2016; Maggio et al., 2011; Mullins et al., 2013; Polanin et al., 2020). Nonetheless, there is still room for improvement in aspects such as indicating the limits of the search, specifying search dates or including the full search strategy. Using the same definition, we found the search reproducible in 37% of the meta-analyses, as opposed to the 22% reported in Koffel and Rethlefsen (2016). In any case, the inclusion of a full reproducible search strategy was modest in the set of meta-analyses reviewed. As recommended in Table 2, and in line with the updated PRISMA 2020 (Page et al., 2021), the full search strategy for all databases consulted, detailing dates, limits, specific terms, and the Boolean connectors should be reported. These details could be reported as additional/supplementary information hosted by the journal or third-party repositories.

The validity of a systematic review partially depends on the reliability of the data extraction process. Coding primary studies requires time, attention to details in a tedious task and multiple choices. Close to one third of the meta-analyses reviewed did not give details on how the study coding process was carried out. In addition, although most of the meta-analyses that reported details of this process carried out double coding, only a third of these reported inter-coder reliability estimates of the coding process. Moreover, missing data is a common problem in evidence synthesis, but only 42% of the meta-analyses reviewed reported any method to deal with missing data. Several methods have been developed to check the robustness of the results to the inclusion of missing data (Mavridis et al., 2014; Pigott, 2019).

Table 2.

Summary of results and recommendations on the key points lacking transparency

Point	Reporting rate	Why is it important?	Recommendations
Completely reproducible electronic search	37% [28%-47%]	Facilitates the evaluation of the comprehensiveness of the review and its update in the same direction.	Always report the full search strategy for ALL databases consulted, detailing dates, limits, specific terms, and the Boolean connectors. For space-saving reasons, it is recommended to report these details as supplementary material hosted by the journal or online repositories.
Specify effect measure formula	15% [9%-24%]	Due to the variety of approaches to define standardized and unstandardized mean differences, specification of the formula used is required to ensure the reproducibility of results.	Always report the specific formula on the paper itself or refer readers to a reference (including the equation number and/or the book/article page where the formula can be found).
Identify the weighting factor	30% [22%-40%]	Although inverse variances are the most popular weighting scheme, other alternatives are available, and the choice can have an impact on the results.	Always specify the weighting factor used. Note that this should only take a few words.

Identify the estimation method of the heterogeneity variance, τ^2	13% [7%-21%]	The between-studies (or heterogeneity) variance is used in random-effects weights and prediction intervals, as well as in the calculation of popular indices in meta-analysis such as I^2 and pseudo-R^2. Many estimators of τ^2 have been proposed, and the resulting estimates often show important discrepancies among estimators.	Always report and justify the estimation method of the heterogeneity variance. The choice should be based on the dataset features along with recommendations from simulation studies under conditions similar to those of the meta-analytic database.
Open availability of statistics used to compute the effect size	30% [21%-39%]	This is the primary raw data used to calculate the effect measures. Availability of this information, along with the effect measure formula, allows the analytic reproducibility of primary effect measures.	Always share ALL coded raw data prior to any data handling in easily computer-readable formats, such as *tsv* or *csv*. To facilitate error checking, add a column indicating the precise location of the coded data in the primary study. Online repositories are very useful for this (OSF, Fighshare, Zenodo, GitHub…), but other options include journal or personal websites.
Interoperability of data sharing format	3% [1%-9%] 3% [1%-9%] 4% [1%-10%] 4% [1%-12%] 7% [2%-22%] 4% [1%-12%]	Significantly increases the efficiency of data reusability through the use of computer-readable and non-proprietary value formats. Avoiding the error-prone process of manual recoding of available data for reproduction or reuse attempts.	Always share data in interoperable formats such as *csv or tsv*. The FAIR principles (Wilkinson et al., 2016) are a useful guideline for best practices in data sharing.

Open availability of analysis script code	1% [0%-5%]	It contains a detailed step-by-step description of the analyses performed. Sharing it is the best way to ensure the analytic reproducibility and to avoid the ambiguities of verbal descriptions.	Always share the analysis script code. Moreau and Gamble (2019) share a very useful script template for carrying out a meta-analysis with R using the metafor (Viechtbauer, 2010) package in their OSF project: https://osf.io/5nk92/. Again, online repositories, own websites or journal hosting are very useful to host the files.

Note: 95% CIs are presented in brackets.

Table 3.

Summary of results and recommendations on different practices related to promoting transparency

Point	Practice rate	Why is it important?	Recommendations
Use of reporting guidelines	30% [20%-40%]	It’s a very helpful tool that facilitates the transparent reporting of all relevant points on the rationale, methods and results of a systematic review or meta-analysis. Furthermore, it standardizes the report, facilitating the readability, assessment and update of the systematic review and/or meta-analysis.	Use well-established, up-to-date reporting guidelines intended for meta-analyses such as: the recently updated PRISMA 2020 (Page et al., 2021); the focused-on reliability generalization meta-analyses REGEMA (Sanchez-Meca et al., 2021); the focused-on non-intervention studies NIRO-SR (Topor et al., 2020), for example.
Pre-registration	19% [12%-17%]	It prevents the result-based bias by stating the main hypotheses, design and analysis plans prior to obtaining the results. Furthermore, it could provide a transparent project timeline, workflow and general decision-making process.	Specialized repositories such as PROSPERO could be helpful since they are tailored to the SR/MA design. General repositories such as OSF could also be helpful as they provide a useful space to store all relevant material related to the project. It’s important to note that a pre-registration protocol does not restrict flexibility. Deviations from the pre-registration protocol are normal and usual, they should simply be reported.

Previous studies examined the reproducibility of primary effect sizes of a set of meta-analyses: Gøtzsche et al., (2007) found problems in 37% of these meta-analyses and, Lakens et al. (2017) found significant problems to reproduce a set of meta-analyses, in part due to the lack of information on how the primary effects sizes were calculated and Maassen et al., (2020) found that the main problems with primary effect sizes reproducibility are often related to the ambiguity in the procedure followed by the meta-analyst. Thus, reporting information concerning the primary effects sizes used and their exact and detailed computation methods is essential to reproduce and update a meta-analysis. However, a poor reporting of detailed primary effect sizes computation method was found in our study. As pointed in Table 2, due to the variety of approaches to compute a common kind of effect measure (e.g., *d* index family), more detailed information on this should be specified. Commonly, general references to handbooks have been found, but the specific computation method used should be specified with a mention to the page(s) where the calculation formula(e) can be found. Furthermore, multiplicity of results in primary studies is a common meta-analysis issue and the way to deal with it could have an impact on the meta-analytic model estimates (Massen et al., 2020; Tendal et al., 2011), but only half of the meta-analyses reviewed reported any method of dealing with it. Also, it is common to find extreme effect sizes in a set of primary studies when carrying out a meta-analysis. Apart from this, the presence of outliers could have an impact on the conclusions, however, only a third of the meta-analyses reviewed dealt with this issue. There are different approaches to handling influential observations such as leave-one-out analyses and Cook’s distances (Viechtbauer, 2010) or graphical examination of heterogeneity using combinatorial meta-analysis (Olkin et al., 2012). Addressing the issue of influential results is a good practice to appraise the robustness of the conclusions derived from the quantitative synthesis.

Regarding synthesis methods, different analytic choices have to be made when a meta-analysis is carried out. As pointed in Table 2, these choices could have an impact on the results (Langan et al., 2015; Sánchez-Meca et al., 2013; Schmidt et al., 2009) and compromise the reproducibility of the meta-analysis and should be reported. However, a lack of transparency was found in the report of relevant information such as the weighting factor used or the estimation method of the heterogeneity variance when a random-effects model was assumed. On the one hand, a comprehensive description of the synthesis methods used in a meta-analysis facilitates the reproducibility, and, on the other hand, it allows the assessment of the robustness of the results when applying different statistical techniques (Steegen et al., 2016). If the meta-analysis is carried out using the R (R Core Team, 2019) package *metafor* (Viechtbauer, 2010) a very helpful function is *reporter()*. This function generates a readable text format output with a draft analysis report based on a previously fitted *rma.uni* object. Such draft may be used as a starting point when writing up the meta-analytic report.

Along with a comprehensive description of the synthesis methods, the availability of open data is the next key aspect that enables the reproducibility of the results as well as checking their robustness. Previous studies found poor ratios of data sharing in primary research in different areas (Alsheikh-Ali et al., 2011; Hardwicke et al., 2020b; Hardwicke et al., 2020c; Hardwicke & Ioannidis, 2018a; Iqbal et al., 2016; Wallach et al., 2018). Despite the majority of the meta-analyses we reviewed reported at least some raw data, most data were shared in the article itself. Indeed, the vast majority of raw shared data was reported in PDF format, hampering reanalysis attempts by different researchers and most likely forcing them to tedious, time-consuming and, and error-prone manual recoding of the data (Bek, 2019; Wilkinson et al., 2016). Only 3 studies shared some raw data in interoperable formats such as CSV files. On other hand, the shared raw data was typically limited to the primary effect sizes computed (as opposed to the raw data reported in the primary studies). Conversely, it was uncommon to find primary raw statistics used to compute the effect sizes, similar to previous studies (Polanin et al., 2020). This is the process where more problems have been found to reproduce the results of a meta-analysis (Gøtzsche et al., 2007; Massen et al., 2020). There is no good reason for a meta-analyst not to share all the coded raw data. We note that, with the exception of individual participant data meta-analysis, the unit of analysis involves summary data from primary studies, hence sharing the meta-analysis database usually entails no ethical concerns. Nowadays, there are many ways for data sharing in interoperable spreadsheet formats, for example: hosted by the journal, online repositories (e.g., OSF, Fighshare, Zenodo), or personal/institutional webpages. In addition to reproducibility concerns, data sharing allows for quick updating of a meta-analysis and the reusability for new scientific purposes. As mentioned in Table 2, the FAIR principles (Wilkinson et al., 2016) are a useful guideline for best practices in data sharing. Meta-analytic data findable, accessible, interoperable and reusable would have a stronger impact and efficiency by decreasing research waste.

Previously, we discussed the relevance of a comprehensive description of synthesis methods to guarantee the reproducibility of the results. However, this form of verbal description is often lacking in detail or contains errors making reproducibility difficult (Hardwicke et al., 2018; Lakens et al., 2017). A better approach to ensure the analytic reproducibility is sharing the analysis script (Hardwicke et al., 2018; Obels et al., 2020), typically in computer code format. Unfortunately, only one meta-analysis shared the analysis script. This result is in line with previous research (Hardwicke et al., 2020b; Hardwicke et al., 2020c; Polanin et al., 2020; Wallach et al., 2018).

Nowadays, there are many options for analysis script sharing, allowing easily reproducibility and detection of potential errors. R (R Core Team, 2019) is a free and open software environment and programming language that, along with RStudio, facilitates the production of easily-shared analysis scripts. As pointed in Table 2, Moreau and Gamble (2020) share a very useful script template for carrying out a meta-analysis with R using the *metafor* (Viechtbauer, 2010) package in their OSF project: https://osf.io/5nk92/.

The prevalence of funding statements found in our meta-review of meta-analyses of psychological interventions was similar to those reported in the broader fields of psychology (Hardwicke et al., 2020b) and biomedical research (Wallach et al., 2018), and higher than social sciences research (Hardwicke et al., 2020c). Regarding competing interests, better ratios to include a statement were found compared to psychology and social sciences research, and similar to biomedical research. Accessibility was fairly adequate compared to biomedical (Wallach et al., 2018) and social sciences (Hardwicke et al., 2020c) research and similar to psychology (Hardwicke et al., 2020b). In any case, there is still room for improvement. Of the 29 meta-analyses for which we could not find any publicly available version, 13 stated that public funding was provided. Public research funders usually have open-access mandates (van Noorden, 2021), which make sense. Green open-access consists of self-archiving a copy of the work in a freely accessible repository (institutional, third-party archive…) or personal webpage and does not entail any extra charge for the authors. Different versions of the manuscript such as pre-print or an author-accepted version, can be stored.

This study has some limitations. First, the time span covered is fairly wide. Thus, the obtained estimates may not capture the changes that have arisen in recent years. Due our focus on a highly specific area of research our primary goal was to capture general transparency and reproducibility-related practices over a wide time span, and then we subsequently attempted to assess possible variations over time using logistic regression models with publication year as a predictor. Therefore, additional research is needed to examine more specific changes over years. Second, our conclusions might not be generalizable beyond the area of clinical psychology. Additional research is needed to address these issues in different meta-analytic contexts. Third, this study was not pre-registered. Although the nature of our analyses is strictly exploratory, there are several benefits of pre-registration for all kinds of studies regardless of their design or aims. Mainly, regarding workflow and the decision-making process transparency. We have attempted to address this gap by openly sharing all relevant material at the different stages of the study. Last, our results do not provide findings on the reproducibility of the meta-analyses reviewed, but on the prevalence of transparency and reproducibility-related practices. The reports were reviewed to assess the availability of necessary information and data to be able to check the reproducibility of a meta-analysis. Further research is needed that specifically addresses the analytic reproducibility of published meta-analyses in different research areas.

Our findings show a relatively better level of transparency and reproducibility-related practices across meta-analyses on the effectiveness of psychological interventions compared to more general fields or research areas. Nevertheless, some gaps were found in key aspects such as: full reproducible search, level of detail on statistical methods, availability and interoperability of relevant raw data and script analysis code sharing. Nowadays, meta-analysis is widely considered as the best source of scientific evidence (e. g., OCEBM Levels of Evidence Working Group, 2011) and therefore meta-analytic results and conclusions often have a strong impact on policy making, social practices, or healthcare judgement. Thus, standards of research quality, transparency, and reproducibility-related practices of meta-analyses need to be high. Tools to help researchers carry out a meta-analysis with the best open practices are available (e.g., Lakens et al., 2016; Moreau & Gamble, 2020), as well as a recent update of the PRISMA statement (Page et al., 2020). We also provide some recommendations in Table 2 which are particularly relevant to researchers carrying out evidence synthesis in the field of clinical psychology. Increasing compliance to these different recommendation sources will improve the strength of the conclusions of a meta-analysis and will allow a more efficient and stronger development of scientific knowledge. These points are particularly relevant in the context of meta-analytic research recognized and understood as a source of evidence synthesis commonly used to guide applied practice. Flawed meta-analytic conclusions could lead to misguided practical applications, particularly harmful in healthcare context. Last, this study provides a baseline for comparison that will allow future studies to assess the impact of recent developments in this field.

References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PloS one*, *6*(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. van, Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119. https://doi.org/10.1002/per.1919

Bek, J. G. (2019). *Bringing order to psychological data: Explorations in a meta-analytical space* [Master's Thesis, Eindhoven university of technology]. https://research.tue.nl/en/studentTheses/bringing-order-to-psychological-data

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N. L., & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology*, *61*(9), 857-865. https://doi.org/10.1016/j.jclinepi.2008.03.004

Biondi-Zoccai, G. (Ed.)(2016). *Umbrella reviews: Evidence Synthesis with overviews of reviews and meta-epidemiologic studies*. Springer. https://doi.10.1007/978-3-319-25655-9

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333-342. https://doi.org/10.1177/1745691614529796

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145-155. https://doi.org/10.1177/2515245919847421

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77-84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PLOS ONE, 13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gøtzsche, P. C., Hróbjartsson, A., Maric, 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

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573. https://doi.org/10.1177/1745691616652873

Hardwicke, T. E., & Ioannidis, J. P. A. (2018a). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE, 13*(8), e0201856. https://doi.org/10.1371/journal.pone.0201856

Hardwicke, T. E., & Ioannidis, J. P. A. (2018b). Mapping the universe of registered reports. *Nature Human Behaviour, 2*(11), 793-796. https://doi.org/10.1038/s41562-018-0444-y

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr A. H., Clayton, E., Yoon, E. J., Tessler, M. H., 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., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020a). Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application, 7*(1), 11-37. https://doi.org/10.1146/annurev-statistics-031219-041104

Hardwicke, T. E., Thibault, R. T., Kosie, J., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2020b). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017). MetaArXiv. https://doi.org/10.31222/osf.io/9sz2y

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020c). 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

Hoyt, W. T., & Del Re, A. C. (2018). Effect size calculation in meta-analyses of psychotherapy outcome research. *Psychotherapy Research, 28*(3), 379-388. https://doi.org/10.1080/10503307.2017.1405171

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. https://doi.org/10.1371/journal.pbio.2005468

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *PLOS Biology, 14*(1), e1002333. https://doi.org/10.1371/journal.pbio.1002333

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association, 112*(517), 1-10. https://doi.org/10.1080/01621459.2016.1240079

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), 142-152. http://dx.doi.org/10.1027/1864-9335/a000178

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

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

Lakens, D, Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3

Lakens, D., Page-Gould, E., van Assen, M., Spellman, B., Schönbrodt, F. D., Hasselman, F., Corker, K., Grange, J., Sharples, A., Cavender, C., Augusteijn, H., Gerger, H., Locher, C., Miller, I., Anvari, F. & Scheel, A. M. (2017). Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report. https://doi.org/10.31222/osf.io/xfbjf

Langan, D., Higgins, J. P., & Simmonds, M. (2015). An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Research synthesis methods*, *6*(2), 195-205. https://doi.org/10.1002/jrsm.1140

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ, 339*. https://doi.org/10.1136/bmj.b2700

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research synthesis methods*, *9*(3), 336-351. https://doi.org/10.1002/jrsm.1310

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

Maassen, E., van Assen, M. A., 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

Mavridis, D., Chaimani, A., Efthimiou, O., Leucht, S., & Salanti, G. (2014). Addressing missing outcome data in meta-analysis. *Evidence-based mental health, 17*(3), 85-89. https://doi.org/10.1136/eb-2014-101900

McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229-229. https://doi.org/10.1126/science.1250475

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

Mullins, M. M., DeLuca, J. B., Crepaz, N., & Lyles, C. M. (2014). Reporting quality of search methods in systematic reviews of HIV behavioral interventions (2000–2010): are the searches clearly explained, systematic and reproducible? *Research Synthesis Methods, 5*(2), 116-130. https://doi.org/10.1002/jrsm.1098

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

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer, 31*(3). https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A.., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences, 23*(10), 815-818. https://doi.org/10.1016/j.tics.2019.07.009

Nuijten, M. B., Assen, M. A. L. M. van, Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172-182. https://doi.org/10.1037/gpr0000034

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229-237. https://doi.org/10.1177%2F2515245920918872

OCEBM Levels of Evidence Working Group (2011). *The Oxford 2011 Levels of Evidence.* *https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence*

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH–a graphical display of study heterogeneity. *Research Synthesis Methods*, *3*(3), 214-223. https://doi.org/10.1002/jrsm.1053

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

Page, M. J., McKenzie, J. E., & Forbes, A. (2013). Many scenarios exist for selective inclusion and reporting of results in randomized trials and systematic reviews. *Journal of clinical epidemiology*, *66*(5), 524-537. https://doi.org/10.1016/j.jclinepi.2012.10.010

Page, M. J., & Moher, D. (2017). Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review. *Systematic reviews*, *6*(1), 1-14. https://doi.org/10.1186/s13643-017-0663-8

Page, M. J., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., & Mulrow, C. D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530. https://doi.org/10.1177/1745691612465253

Pigott, T. D. (2019). Missing data in Meta-Analysis. In H. Cooper, L. V. Hedges & J. C. (Eds.) Valentine. *The handbook of research synthesis and meta-analysis* 3rd ed. (pp. 367-382). Russell Sage Foundation.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research, 90*(1), 24-46. https://doi.org/10.3102/0034654319877153

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

Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature*, *569*(7756), 445-447. https://doi.org/10.1038/d41586-019-01506-x

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods*, *50*(5), 2057-2073. https://doi.org/10.3758/s13428-017-0973-8

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *The British Journal of Mathematical and Statistical Psychology, 66*(3), 402-425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibañez, C. & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*. Advanced online publication. https://doi.org/10.1002/jrsm.1487

Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*(1), 32. http://dx.doi.org/10.1037/arc0000029

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *The British Journal of Mathematical and Statistical Psychology, 62*(1), 97-128. https://doi.org/10.1348/000711007X255327

Signorell, A. et al. (2020). DescTools: Tools for descriptive statistics. R package version 0.99.38. https://CRAN.R-project.org/package=DescTools

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Sison, C. P., & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association, 90*(429), 366-369. https://doi.org/10.2307/2291162

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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712. https://doi.org/10.1177%2F1745691616658637

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

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*. 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*. https://doi.org/10.1136/bmj.d4829

Topor, M., Pickering, J. S., Barbosa Mendes, A., Bishop, D. V. M., Büttner, F. C., Elsherif, M. M., … Westwood, S. J. (2020, December 14). An integrative framework for planning and conducting Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR). *MetaArXiv*. https://doi.org/10.31222/osf.io/8gu5z

van Assen, 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

Van Noorden, R. (2021). Do you obey public-access mandates? Google Scholar is watching. *Nature*. https://doi.org/10.1038/d41586-021-00873-8

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48. http://dx.doi.org/10.18637/jss.v036.i03

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 819–824. https://doi.org/10.1145/2110363.2110464

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

Westgate, M. J. (2019). revtools: An R package to support article screening for evidence synthesis. *Research Synthesis Methods, 10*(4), 606-614. https://doi.org/10.1002/jrsm.1374

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical Inference. *Journal of the American Statistical Association, 22*(158), 209-212. https://doi.org/10.1080/01621459.1927.10502953

Chapter 3

Reproducibility of published meta-analyses on clinical psychological interventions[3]

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.

[3] **Lopez-Nicolas, R**., Lakens, D., López-López, J., Rubio-Aparicio, M.., Sandoval-Lentisco, A., López-Ibáñez, C., Blázquez-Rincón, D. & Sánchez-Meca, J. (in press). Reproducibility of published meta-analyses on clinical psychological interventions. *Advances in Methods and Practices in Psychological Science.*

3.1 Introduction

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.

3.1.1 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.

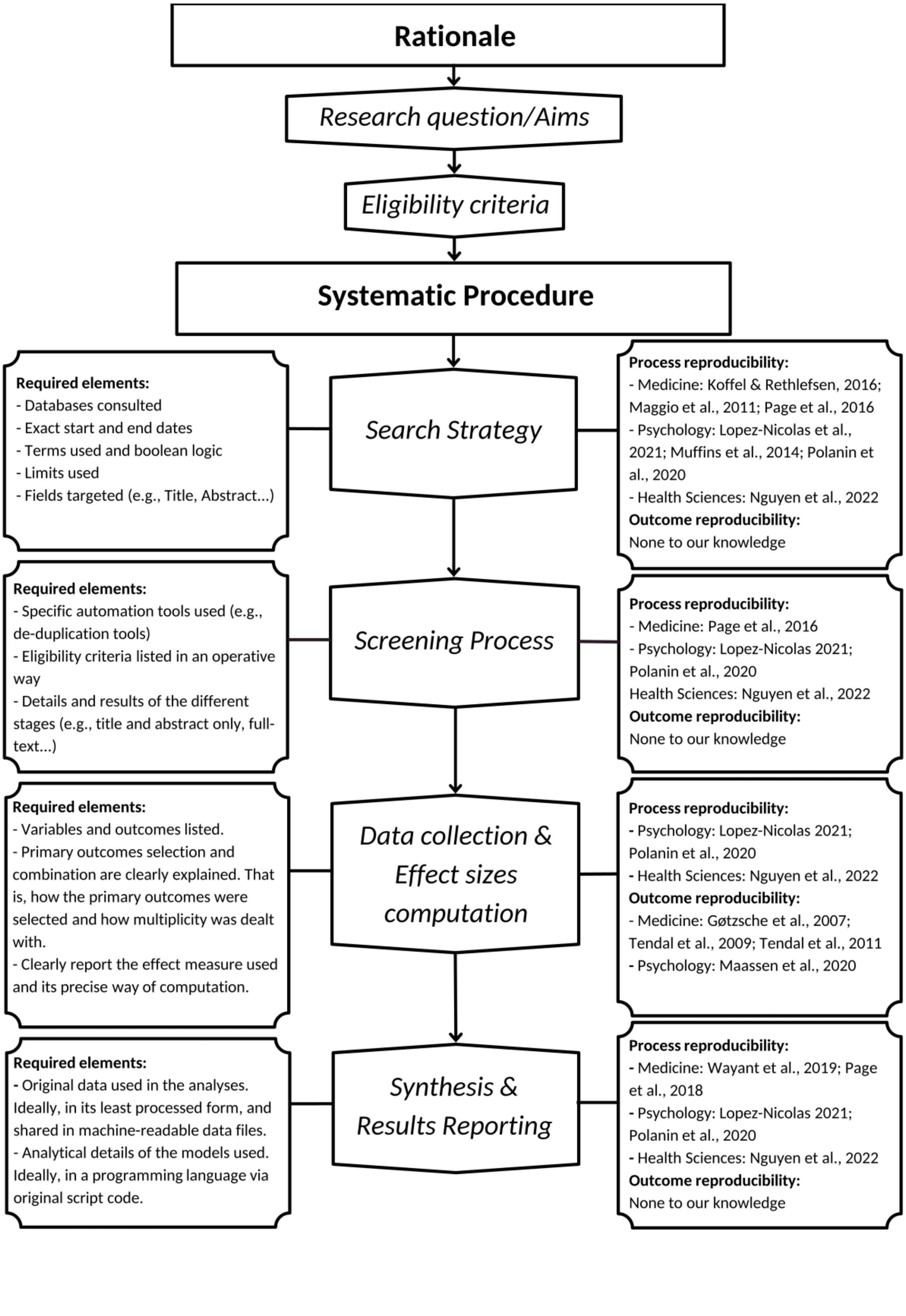

Rationale
Research question/Aims
Eligibility criteria
Systematic Procedure
Search Strategy
Required elements:
- Databases consulted
- Exact start and end dates
- Terms used and boolean logic
- Limits used
- Fields targeted (e.g., Title, Abstract...)
Process reproducibility:
- Medicine: Koffel & Rethlefsen, 2016; Maggio et al., 2011; Page et al., 2016
- Psychology: Lopez-Nicolas et al., 2021; Muffins et al., 2014; Polanin et al., 2020
- Health Sciences: Nguyen et al., 2022
Outcome reproducibility:
None to our knowledge
Screening Process
Required elements:
- Specific automation tools used (e.g., de-duplication tools)
- Eligibility criteria listed in an operative way
- Details and results of the different stages (e.g., title and abstract only, full-text...)
Process reproducibility:
- Medicine: Page et al., 2016
- Psychology: Lopez-Nicolas 2021; Polanin et al., 2020
Health Sciences: Nguyen et al., 2022
Outcome reproducibility:
None to our knowledge
Data collection & Effect sizes computation
Required elements:
- Variables and outcomes listed.
- Primary outcomes selection and combination are clearly explained. That is, how the primary outcomes were selected and how multiplicity was dealt with.
- Clearly report the effect measure used and its precise way of computation.
Process reproducibility:
- Psychology: Lopez-Nicolas 2021; Polanin et al., 2020
- Health Sciences: Nguyen et al., 2022
Outcome reproducibility:
- Medicine: Gøtzsche et al., 2007; Tendal et al., 2009; Tendal et al., 2011
- Psychology: Maassen et al., 2020
Synthesis & Results Reporting
Required elements:
- Original data used in the analyses. Ideally, in its least processed form, and shared in machine-readable data files.
- Analytical details of the models used. Ideally, in a programming language via original script code.
Process reproducibility:
- Medicine: Wayant et al., 2019; Page et al., 2018
- Psychology: Lopez-Nicolas 2021; Polanin et al., 2020
- Health Sciences: Nguyen et al., 2022
Outcome reproducibility:
None to our knowledge

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.

3.2 Method

3.2.1 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 maximize 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.

3.2.2 Retrieval of primary data

In order to be able to reproduce meta-analyses of aggregate data, primary-level[4] 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 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):

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

finally, we sought for the primary effects sizes and their confidence limits, from which the standard errors were approximated as follows:

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

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.

3.2.3 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/42r3p) and a Codebook describing these variables is available at: https://osf.io/vrty7.

3.2.4 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 Appendix 3A.

3.2.5 Reproducibility outcomes

Each meta-analysis was labelled using the following two-level[5] 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

[5] 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.

where we were unable to access the primary data neither through direct extraction nor upon request[6]. 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/42r3p). 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%.[7] 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 synthesis methods such as heterogeneity statistics.

3.2.6 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

[6] 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.

[7] A sensitivity analysis using other possible criteria is reported in the Appendix 3B.

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 the Appendix 3C.

3.3 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)[8]. 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.

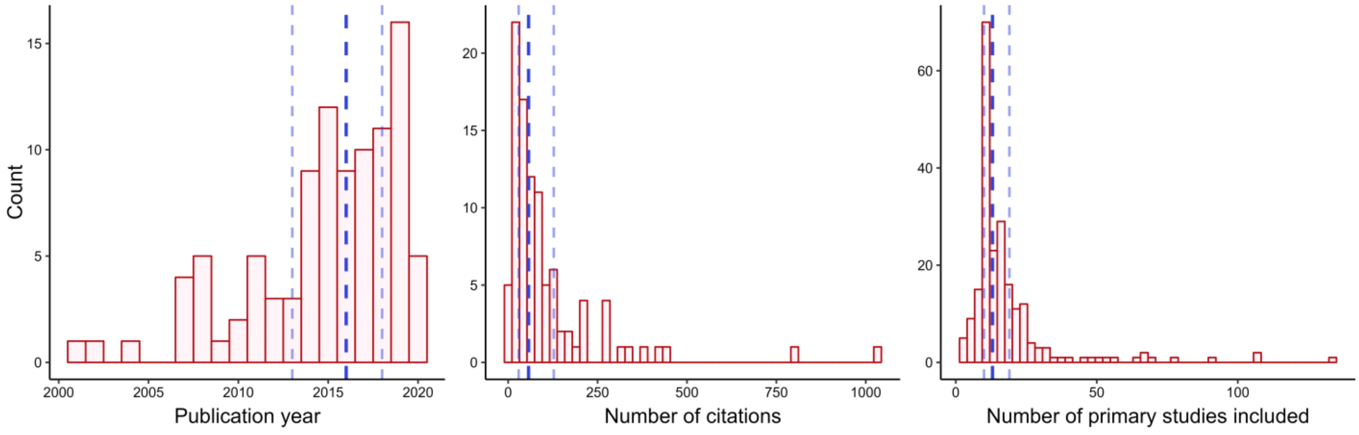


Figure 2. Distribution of (a) the number of primary studies included in each of the meta-analyses; (b) the publication year of the included papers; (c) citation count of the included papers. Vertical blue dotted lines represent the first quartile, median, and third quartile, respectively.

[8] 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.

3.3.1 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 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 Appendix 3C.

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 78 meta-analyses from 25 different papers by emailing the corresponding authors, data was only retrieved for 7 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.

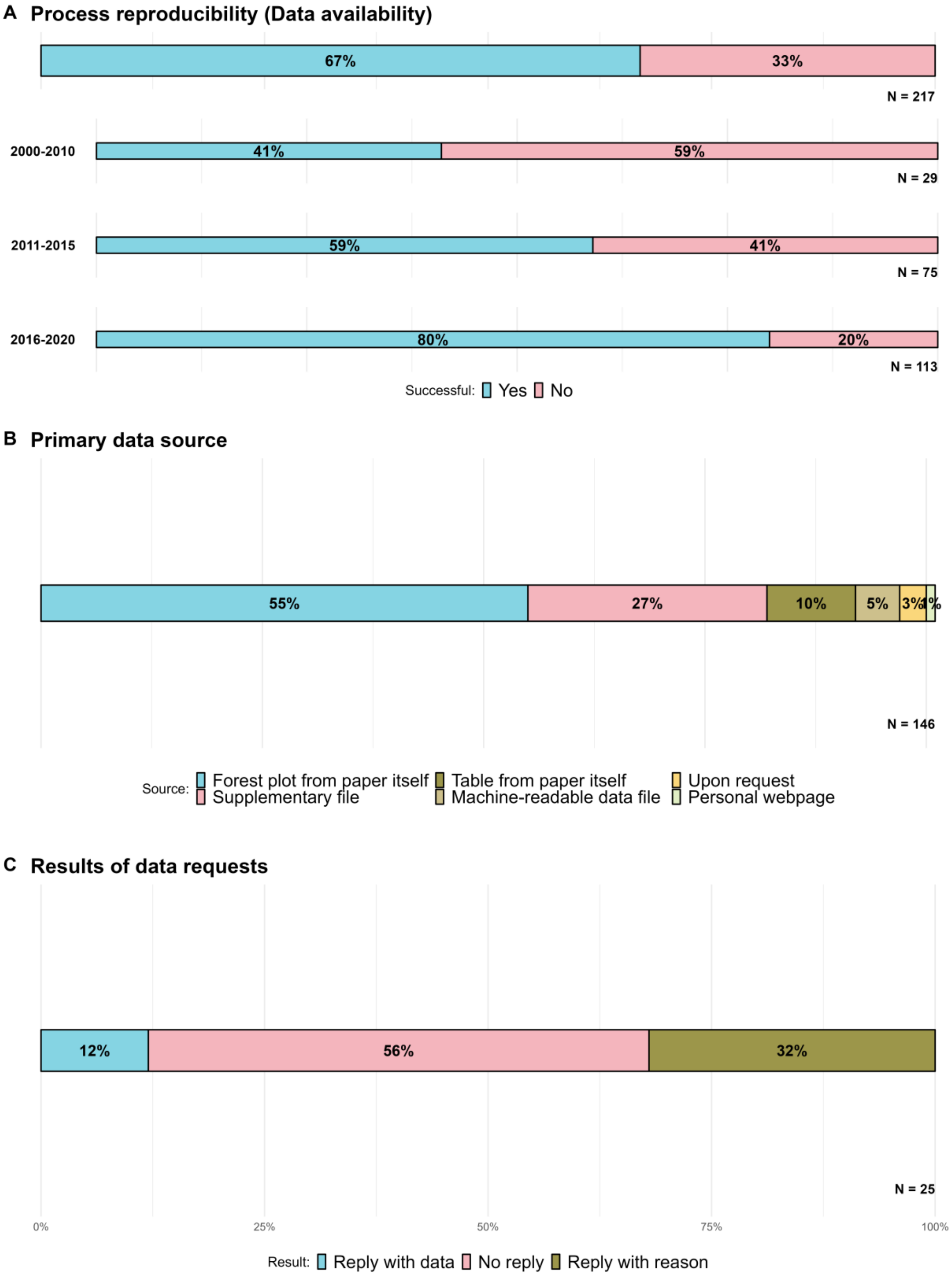


Figure 3. Percentage of (a) process-reproducible meta-analyses; (b) different types of sources of original data; (c) data request results.

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
Proprietary 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

3.3.2 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 7 meta-analyses (from three different papers) of the 146 meta-analyses labelled as process reproducible (5%), 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.

3.3.3 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.

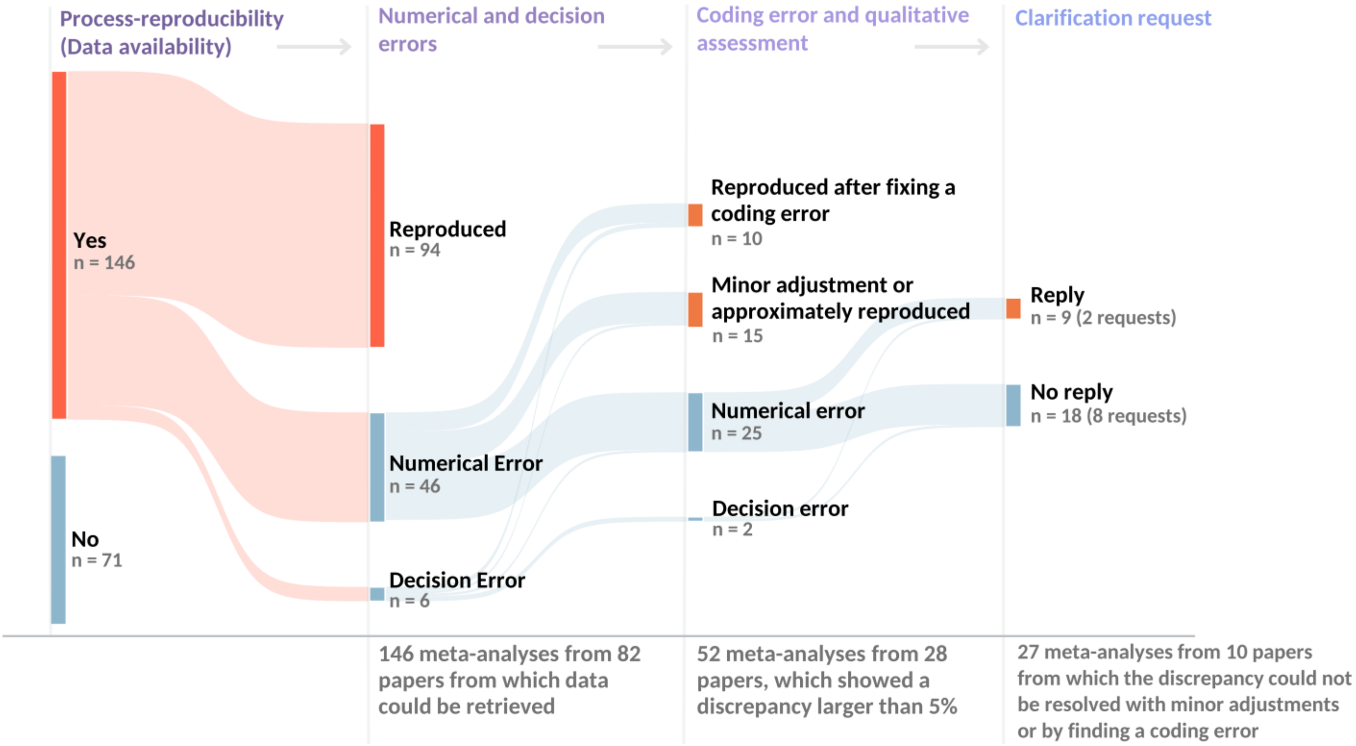

Process-reproducibility (Data availability)
Numerical and decision errors
Coding error and qualitative assessment
Clarification request
Yes
n = 146
No
n = 71
Reproduced
n = 94
Numerical Error
n = 46
Decision Error
n = 6
Reproduced after fixing a coding error
n = 10
Minor adjustment or approximately reproduced
n = 15
Numerical error
n = 25
Decision error
n = 2
Reply
n = 9 (2 requests)
No reply
n = 18 (8 requests)
146 meta-analyses from 82 papers from which data could be retrieved
52 meta-analyses from 28 papers, which showed a discrepancy larger than 5%
27 meta-analyses from 10 papers from which the discrepancy could not be resolved with minor adjustments or by finding a coding error

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[9]. 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.

[9] Full details in Appendix 3D.

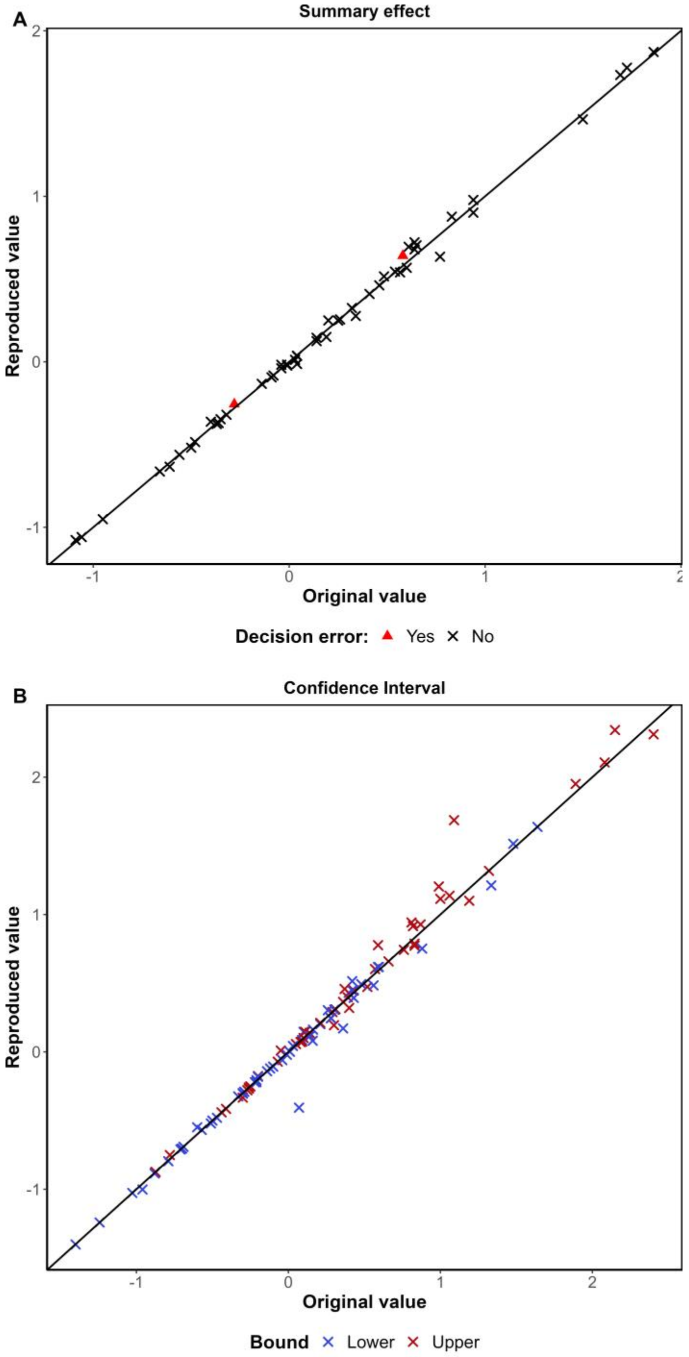


Figure 5. Scatterplot displaying the reproduced values as a function of the original values classified by whether or not decision error was found. Only the results of the 52 meta-analyses with a discrepancy of more than 5% identified in the first stage are displayed, but with the corrections made in 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.

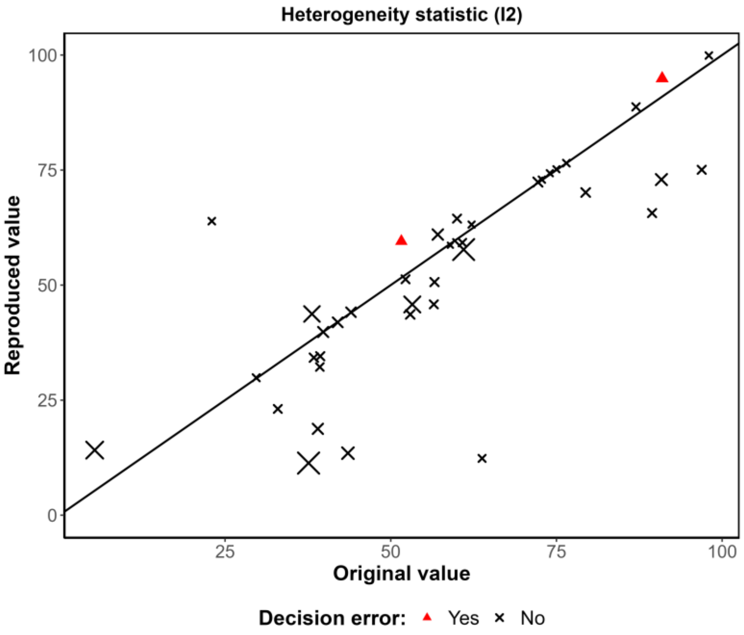


Figure 6. Scatterplot displaying the reproduced values as a function of the original values classified by whether or not decision error was found. Only the results of the 52 meta-analyses with a discrepancy of more than 5% identified in the first stage are displayed, but with the corrections made in the second stage. The values displayed are I2 heterogeneity statistics. The size of the crosses is a function of the discrepancy in the summary effect.

3.3.4 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 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.

3.3.5 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[10] 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.

[10] 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.

3.4 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 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.

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). *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. *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/

Chapter 4

Statistical power of random-effects meta-analyses on clinical psychological interventions

Abstract

Underpowered studies are ubiquitous in psychology and related disciplines. Meta-analysis can help alleviate this problem, increasing the statistical power by combining the results of a set of primary studies. However, this is not necessarily true when we use a random-effects model, which is currently the predominant approach when carrying out meta-analyses. In this study, we examined the statistical power of a sample of 141 meta-analyses on the effectiveness of clinical psychological interventions. Additionally, we compared the estimated statistical power of these meta-analyses with the power of the individual studies that comprised them. To do so, we used different analytical approaches and a Monte Carlo approach. The statistical power of random-effects meta-analyses was computed under different values of the true effect size and levels of heterogeneity. Our results show that under certain scenarios, the hypothesis test of the null-hypothesis of no average effect is underpowered, even showing a lower statistical power than the average or maximum statistical power of included primary studies. Overall, these scenarios were characterised by high heterogeneity and a low number of included studies. While this pattern is expected, our findings show the steepness of this drop in statistical power. These results are discussed in light of the statistical and conceptual basis of random-effects meta-analysis.

4.1 Introduction

Underpowered studies are ubiquitous in psychology and related disciplines such as neuroscience, biomedical science, or the social sciences (Arel-Bundock et al., 2022; de Vries et al., 2022; Dumas-Mallet et al., 2017; Flint et al., 2015; Szucs & Ioannidis, 2017). Meta-analysis can help alleviate this problem and, in fact, increasing the statistical power compared to the primary studies that are included in the meta-analysis- This is one of the most frequently cited motivations to perform a meta-analysis (Cohn & Becker, 2003; Guyatt et al., 2008; Jackson & Turner, 2017; Valentine et al., 2010).

Indeed, this idea always holds true when we assume a fixed-effect model—the more studies we include in the meta-analysis, the smaller the sampling variance and, hence, the larger the power (Cohn & Becker, 2003; Hedges & Pigott, 2001; Jackson & Turner, 2017). However, this is not necessarily true when we use a random-effects model since this model includes one variance component more than a fixed-effect model, the between-study variance (Konstantopoulos & Hedges, 2019). While the fixed-effect model assumes that there is a single a common parametric effect for all estimates of the included studies—i.e., $\theta_1 = \cdots = \theta_k = \theta$, such that $y_i \sim N(\theta, \sigma_i^2)$, with y_i being the effect size estimate of the *ith* study, θ the true effect size, σ_i^2 the sampling variance of the *ith* study, and k the number of included primary studies— the standard random-effects model assumes a normal distribution of parametric effects $\theta_i \sim N(\mu_\theta, \tau^2)$, such that $y_i \sim N(\mu_\theta, \tau^2 + \sigma_i^2)$, with μ_θ being the expected value across studies, and τ^2 the between-study variance component. Therefore, under the fixed effect model the sampling variance of the meta-analytic summary effect size is given by $v = (\sum_{i=1}^{k} 1/\sigma_i^2)^{-1}$, whereas under the random-effects model this sampling variance is given by $v^* = (\sum_{i=1}^{k} 1/(\sigma_i^2 + \tau^2))^{-1}$.

This issue is especially relevant considering that, even though meta-analyses using fixed-effect models were frequent not long ago (Cafri et al., 2010; Field & Gillett, 2010; Hunter & Schmidt, 2000), assuming a random-effects model is currently the predominant approach, as its assumptions are considered more realistic (Aguinis et al., 2010; Borenstein, 2019; Tipton et al., 2019). Moreover, power analysis in random-effects meta-analysis might not be as straightforward as in fixed-effect meta-analyses and different approaches have been proposed on how to estimate power. Next, a brief overview of these proposals (see Jackson & Turner (2017) for further details) is provided.

4.1.1 Statistical power of random-effects meta-analysis

Hedges & Pigott (2001) analytic approach

Following the standard statistical testing procedure in random-effects meta-analysis, the null hypothesis is rejected at level α, if:

$$\frac{|\hat{\mu}_\theta|}{\sqrt{v *}} > z_{1-\alpha/2} \ (1),$$

where $\hat{\mu}_\theta$ is the summary meta-analytic effect size estimation, and $z_{1-\alpha/2}$ is the critical value of the standard normal distribution at the α level.

Therefore, according to Hedges and Pigott (2001), the statistical power formula for a two-tailed test is given by:

$$\mathrm{Power}_{\mathrm{MA}} = 1 - \phi\left(\mathrm{z}_{1-\alpha/2} - \frac{\mu_\theta}{\sqrt{v^*}}\right) + \phi\left(-\mathrm{z}_{1-\alpha/2} - \frac{\mu_\theta}{\sqrt{v^*}}\right) \ (2),$$

where $\phi(\cdot)$ is the cumulative standard normal distribution function, and μ_θ is the true effect size.

Extension based on Knapp and Hartung adjustment

The Knapp-Hartung method (Hartung, 1999; Hartung & Knapp, 2001) is a refined test for the summary meta-analytic effect size. Basically, it applies an adjustment to the sampling variance through a scaling factor and uses a *t-distribution* with *k-1* degrees of freedom for making inferences, attempting to take into account the uncertainty in between-study variance estimation. The scaling factor is given by:

$$H^{*2} = \frac{\sum w_i (y_i - \hat{\mu}_\theta)^2}{k - 1} \ (3),$$

where w_i is the $i = 1 \cdots k \ \ 1/(\sigma_i^2 + \tau^2)$.

Therefore, the adjusted sampling variance is given by:

$$v_{KNHA}^* = v^* H^{*2} \ (4)$$

This method has shown better performance in terms of controlling the Type-I rate than the standard method in several simulation studies (Hartung and Knapp, 2001; InHout et al., 2014; Röver et al., 2015; Sánchez-Meca & Marín-Martínez, 2008). However, in some scenarios this method may present some problems. For instance, although one of its primary strengths is its conservative nature relative to the standard method, it can become more liberal in certain circumstances (Wiksten et al., 2016). This occurs when the scaling factor is less than 1. Given its intention and the nature of the model, this issue can be considered as an undesirable property (Jackson et al., 2017). In this sense, an *ad hoc* modification has been proposed (Knapp & Hartung, 2003) which constrains the scaling factor such that $H^{*2} \geq 1$.

Therefore, following this method, the null-hypothesis is rejected at α level, if:

$$\frac{|\hat{\mu}_\theta|}{\sqrt{v^*_{KNHA}}} > t_{k-1,\ 1-\alpha/2} \text{ (5),}$$

where $t_{v,\gamma}$ denotes the γ-quantile of the *t-distribution* with v degrees of freedom.

In this vein, the former analytical approach by Hedges and Pigott (2001) is extended to this proposal as follows:

$$Power_{MA2} = 1 - T_{k-1}\left(\mathrm{t}_{\mathrm{k}-1,\ 1-\alpha/2}\left|\frac{\mu_\theta}{\sqrt{v^*_{KNHA}}}\right.\right) + T_{k-1}\left(-\mathrm{t}_{k-1,\ 1-\ \alpha/2}\left|\frac{\mu_\theta}{\sqrt{v^*_{KNHA}}}\right.\right) \text{ (6),}$$

where $\mathrm{T}_{df}(\cdot\ |\lambda)$ is the cumulative function of non-central t-distribution with *df* degrees of freedom and λ noncentrality parameter, and $t_{\mathrm{df},\ 1\text{-}\alpha/2}$ is the critical value of the central t-distribution with *df* degrees of freedom at α level.

Jackson and Turner (2017) proposal

While the previous method was proposed to address the uncertainty in the between-study variance component, it still treats it as fixed and known, despite it being an estimated component. Jackson and Turner (2017) proposed an alternative analytical method for computing statistical power in random-effect meta-analysis by deriving the cumulative distribution function of the test statistic *T* as follows:

$$\begin{aligned} P(T \leq t) = \Gamma_1\left(\frac{k-1}{2}, \frac{(1-I^2)(k-1)}{2}\right)\phi\left(\left(t - \mu_\theta\sqrt{k}\sigma\right)\sqrt{1-I^2}\right) \\ +2(k-1)\int_{\sqrt{1-I^2}}^{\infty} x\phi\left(tx - \mu_\theta\sqrt{k}\sigma\sqrt{1-I^2}\right)\chi^2_{k-1}\left((k-1)x^2\right)\mathrm{d}x, \end{aligned} \text{ (7),}$$

where I^2 is the heterogeneity statistic by Higgins and Thompson (2002), defined as $I^2 = \frac{\tau^2}{\sigma^2 + \tau^2}$, where σ^2 is the typical within-study variance, and $\chi^2_{k-1}(\cdot)$ is the probability density function of the χ^2 distribution with *(k − 1)* degrees of freedom. See Jackson and Turner (2017) for full details.

Then the statistical power of random-effect meta-analysis is given by:

$$Power_{MA} = 1 - P\left(T \leq z_{1-\alpha/2}\right) + P\left(T \leq -z_{1-\alpha/2}\right) \text{ (8)}$$

This approach has the advantage of taking into account the uncertainty of between-study variance by not constraining the parameter.

Monte Carlo approach

The previous approach assumes that all the primary studies in a meta-analysis are of the same size, for instance by taking the typical within-study variance (Higgins & Thompson, 2002) or the average of a set of within-studies variance. Instead, a Monte Carlo approach can be employed. This involves simulating *n* datasets with *k* values drawn from a standard normal distribution such that: $X_i \sim N(\mu_\theta, \sigma_i^2 + \tau^2), i = 1, \cdots k,$ for a true μ_θ and τ^2 and a set of within-study variances. Subsequently, the summary meta-analytic effect size is computed for each dataset under a random-effects model. Finally, the standard statistical test presented in equation (1), or the adjusted one presented in equations (5), is applied, and the statistical power is determined as the proportion of statistically significant cases.

Yet, research that provides an overview of the actual power of meta-analyses is scarce, and the available studies focus instead on the power of the primary studies that are included in the meta-analyses. Nonetheless, one important finding of these studies is that, consistently across different disciplines, meta-analyses mainly comprise studies that are underpowered—i.e., less power than .80—to detect small and medium effect sizes (de Vries et al., 2011; Dumas-Mallet et al., 2017; Nuitjen et al., 2020; Parish et al., 2021). Studies that have surveyed the power at a meta-analytic level come, in its majority, from the field of medicine and show that, often, meta-analyses are neither adequately powered to detect small or even medium effect sizes (Carvalho et al., 2020; Jia et al., 2021; Turner et al., 2013). Nonetheless, these findings contrast with the results of Cafri et al. (2010), who examined a set of meta-analyses published in *Psychological Bulletin* and found a median meta-analytic power of .99 to detect the mean effect size estimated by each meta-analysis, or the results of Niemeyer et al., (2022), who assessed PTSD psychotherapy meta-analyses and found that the majority had a power higher than .80 to also detect their synthesized effect size. On the other hand, Jackson & Turner (2017) compared the meta-analytical power and the average power of the primary studies of 1991 Cochrane reviews and found that, when meta-analyses had more than five primary studies, >80% of the meta-analyses had greater power than primary studies. However, this proportion was much lower when meta-analyses consisted of only two or three primary studies, where only about 50% and 65% of the meta-analyses, respectively, achieved greater power. To our knowledge, there is no study that has assessed the meta-analytic power of a large sample of meta-analyses of clinical psychological interventions and compared it with the power of the primary studies that are included in those meta-analyses.

4.1.2 Purpose

The first objective of this study was to provide an overview of the power of a random sample of meta-analyses of clinical psychological interventions. For doing so, we used the different methods available to estimate power in random-effect meta-analyses, enabling a comparison of the results obtained from each of these methods. Second, the estimated power of these meta-analyses was contrasted with the power of the individual studies that constituted the meta-analyses. This way, we could evaluate the increase of statistical power and identify cases where the power of the meta-analysis was actually lower than the power of the primary studies. In those cases, we explored the characteristics of those meta-analyses.

4.2 Method

4.2.1 Identification and selection of meta-analyses

In a previous study a pool of 664 meta-analytic reports on clinical psychological interventions published between 2000 and 2020 were identified through a systematic electronic search (López-Nicolás et al., 2022). Of these reports, 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/. From these reports, 217 independent pairwise meta-analysis of aggregate data were selected, and the primary data on which meta-analytic models were fitted could be retrieved for 146 of them in another previous study (Lopez-Nicolas et al., 2023). In this study, 141 of those datasets were included, all of them use standardized mean difference as measure of effect.

4.2.2 Statistical power computation

The statistical power of random-effects meta-analyses was computed using five different procedures: the standard analytical approach by Hedges and Pigott (2001), as given in equation (2); the extension of this approach based on the Knapp-Hartung (Hartung, 1999; Hartung & Knapp, 2001) adjustment for meta-analytic statistical testing, as given in equation (6); the analytical procedure by Jackson and Turner (2017), as given in equations (7) and (8); and a Monte Carlo procedure for both standard testing, as given in equation (1), and Knapp-Hartung adjusted testing, as given in equation (5), by simulating 10,000 datasets for each meta-analysis. For the computations with Knapp and Hartung adjustment, the *ad hoc* modification was employed. Furthermore, the statistical power of individual studies that contribute to a random-effects meta-analysis, as derived by Jackson and Turner (2017), is given by:

$$Power_{IND} = 1 - \Phi\left((z_{1-a/2}\sigma_i + \mu_\theta) / \sqrt{\sigma_i^2 + \tau^2} \right) \\ + \Phi\left((-z_{1-a/2}\sigma_i + \mu_\theta) / \sqrt{\sigma_i^2 + \tau^2} \right) (9)$$

Subsequently, the average, median and maximum of the statistical power power were computed for each set of individual studies included in each meta-analysis. A significance level $\alpha = 0.05$ was used in all the power computations.

4.2.3 Conditions

The statistical power computations were carried out under different conditions. Firstly, three different values of the true effect size were determined —corresponding to the rule-of-thumb by Cohen (1969) for small, medium and large effect size— $\mu_{\theta S} = 0.2$; $\mu_{\theta M} = 0.5$; $\mu_{\theta L} = 0.8$. On the other hand, different scenarios of true heterogeneity were considered. Specifically, the first-quartile, median, and third-quartile of the estimated I^2 values in the sample of meta-analyses were established—$I^2 = 40.4$; $I^2 = 62.6$; $I^2 = 80.4$, respectively. These values were estimated by fitting random-effects models on each included meta-analytic dataset, using the restricted maximum likelihood estimator for between-studies variance. Furthermore, statistical power was computed by using the estimated values of μ_θ and τ^2 of each meta-analysis as true values.

4.3 Results

A total of 141 meta-analyses were included in the study, with an average of 16.87 primary studies per meta-analysis (SD = 16.15; median = 12; interquartile range = 10-18; range = 3-134). Figure 1 illustrates the distribution of the number of primary studies among the meta-analyses included in our sample.

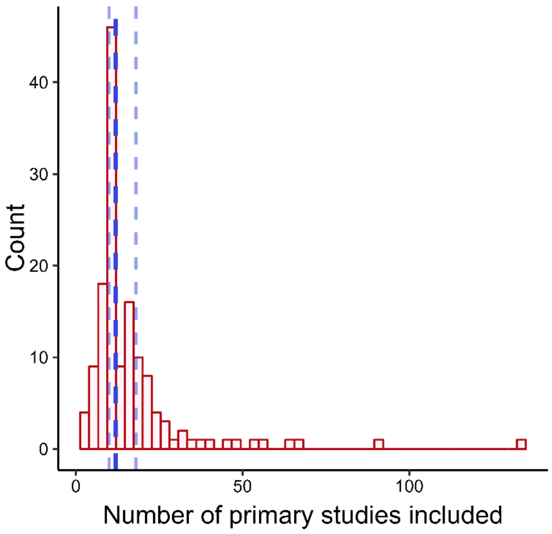


Figure 1. Distribution of the number of primary studies included in each of the meta-analyses. Vertical blue dotted lines represent the first quartile, median, and third quartile, respectively.

4.3.1 Overview of statistical power of meta-analyses

Figure 2 presents a set of boxplots displaying the statistical power of each meta-analysis under the different conditions as result of each method. Taking the results of the Monte Carlo approach as reference, both the Hedges and Pigott (2001) and Jackson and Turner (2017) analytical approaches showed a close alignment with the actual power across various conditions of effect size magnitudes and heterogeneity levels, just showing a slight overestimation in high-powered meta-analyses and a slight underestimation in low-powered meta-analyses. However, the analytical approach incorporating the Knapp and Hartung adjustment appears to systematically underestimate the statistical power, taking Monte Carlo results for Knapp and Hartung testing as reference.

Based on the results from Monte Carlo approach, Tables 1 and 2 show the percentage of included meta-analyses which achieved a statistical power below 80% and below 50% across the different conditions. Table 1 presents the results for standard testing, and Table 2 presents the results for adjusted testing using the Knapp and Hartung testing. The percentages are provided for the total sample, the subset of meta-analyses with less than 10 primary studies and the subset of meta-analyses with 10 or more primary studies.

For both statistical testing approaches, all meta-analyses with less than 10 primary studies showed a statistical power below 80% for detecting an effect size of 0.2 across all heterogeneity scenarios. In fact, the statistical power of all those meta-analyses fell below 50% in the high-heterogeneity scenario for standard testing, and in the medium- and high-heterogeneity scenarios for Knapp and Hartung testing. Overall, the included meta-analyses showed a poor statistical power for detecting an effect size of 0.2 (see also Figure 3, row 1).

Conversely, for effect sizes of medium and large magnitude (0.5 and 0.8), the statistical power was generally more adequate for the majority of meta-analyses, except in cases where there was large heterogeneity and included primary studies were less than 10 (see Table 1 and 2). For instance, 81% and 87% (statistical testing and adjusted testing, respectively) of meta-analyses with less than 10 primary studies showed a statistical power below 80% for an effect size of 0.5 and high heterogeneity.

4.3.2 Statistical power of primary studies included in meta-analyses versus power of the meta-analysis

Tables 3 and 4 show the percentage of cases where the average, median, or maximum statistical power of primary studies that constitute the meta-analysis was larger than the power of meta-analysis itself, based on Monte Carlo approach results, for standard testing, and for adjusted testing using the Knapp and Hartung, respectively. The percentages are provided for the total sample, the subset of meta-analyses with less than 10 primary studies and the subset of meta-analyses with 10 or more primary studies.

Overall, in the low-heterogeneity scenario, meta-analyses generally exhibited higher statistical power than the mean, median, and maximum power of the contributing primary studies. However, as heterogeneity increases, particularly for meta-analyses with less than 10 primary studies, the situation shifts. Specifically, in 94% and 100% (standard testing and Knapp and Hartung testing, respectively) of the cases that included less than 10 primary studies, the average statistical power of included primary studies was greater than the power of meta-analysis itself for a true effect of 0.2 and high heterogeneity. Furthermore, in 65% and 84% (for both testing) of the cases that included less than 10 primary studies, the maximum power of included primary studies was greater than the power of meta-analysis itself for a true effect of 0.5 and high heterogeneity. Moreover, in 48% and 84% (for both testing) of the cases with less than 10 primary studies, the average statistical power of included primary studies was greater than the meta-analytic power for a true effect of 0.2 and medium heterogeneity, being 94% and 100% if the maximum is taken instead of the average (see Table 3 and 4 for full results).

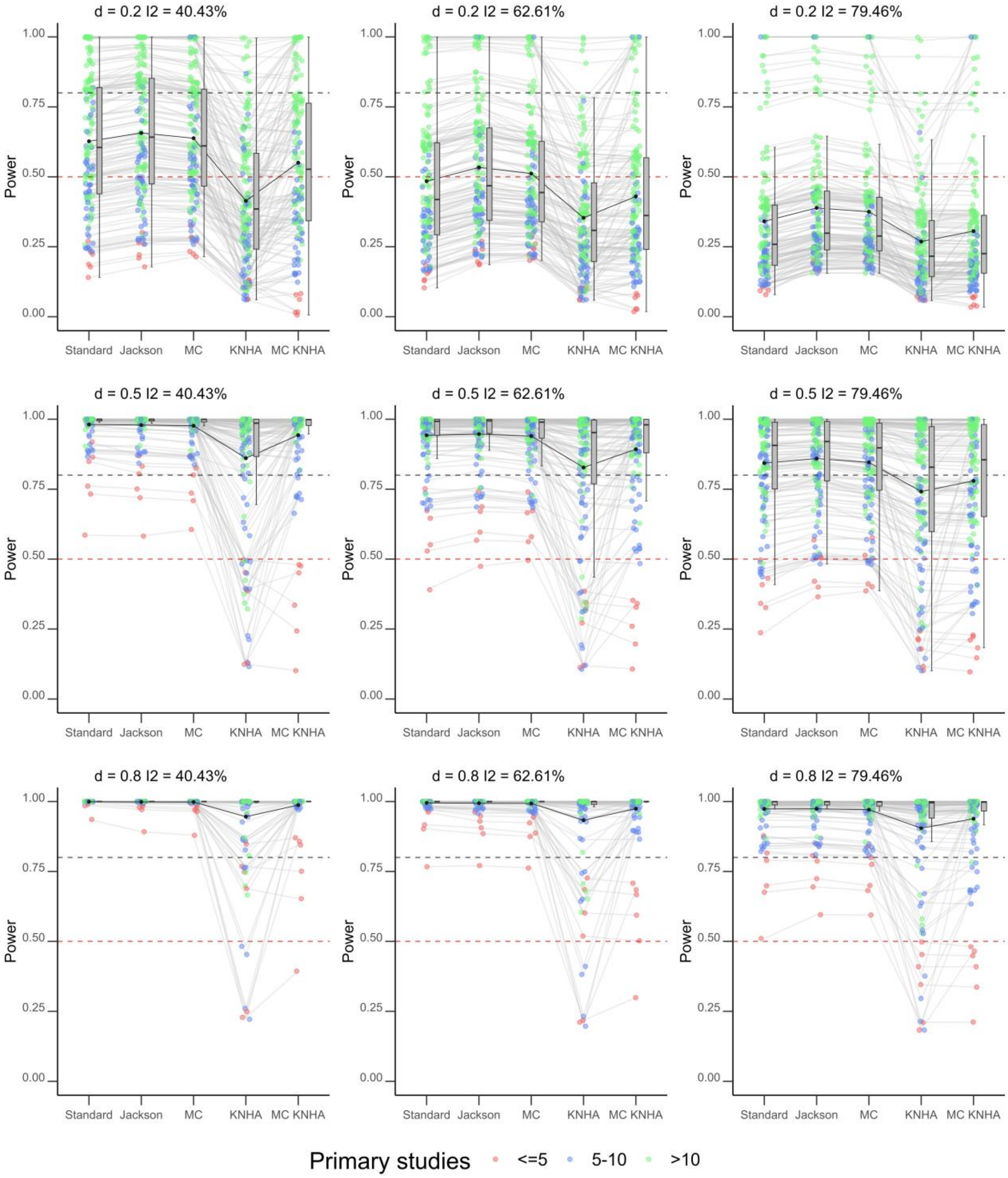


Figure 2. Boxplots displaying the statistical power of included meta-analyses across the different conditions. Black dotted line indicates statistical power of 80% and red dotted line indicates statistical power at 50%.

Table 1. Percentage of included meta-analyses which showed a statistical power below 80% and below 50% across the different conditions for standard testing.

		Statistical power below 80%				**Statistical power below 50%**			
	k/μ_θ	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***
I^2= 40.43	***All***	70%	2%	0%		31%	0%	0%	
	<10	100%	10%	0%		81%	0%	0%	
	≥10	61%	0%	0%		17%	0%	0%	
I^2= 62.61	***All***	87%	11%	1%		57%	1%	0%	
	<10	100%	45%	3%		87%	3%	0%	
	≥10	83%	2%	0%		47%	0%	0%	
I^2= 79.46	***All***	91%	31%	4%		83%	6%	0%	
	<10	100%	81%	16%		100%	26%	0%	
	≥10	89%	17%	0%		78%	1%	0%	
Post-hoc	***All***				29%				19%
	<10				35%				23%
	≥10				27%				18%

Table 2. Percentage of included meta-analyses which showed a statistical power below 80% and below 50% across the different conditions for KNHA testing.

		Below 80%				**Below 50%**			
	k/μ_θ	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***
I^2= 40.43	***All***	82%	9%	2%		48%	4%	1%	
	<10	100%	39%	10%		87%	19%	3%	
	≥10	76%	1%	0%		36%	0%	0%	
I^2= 62.61	***All***	89%	18%	4%		65%	5%	1%	
	<10	100%	71%	19%		100%	23%	3%	
	≥10	86%	4%	0%		55%	0%	0%	
I^2= 79.46	***All***	93%	45%	11%		86%	14%	4%	
	<10	100%	87%	45%		100%	58%	19%	
	≥10	91%	34%	2%		82%	2%	0%	
Post-hoc	***All***				33%				20%
	<10				48%				26%
	≥10				29%				18%

Table 3. Percentage of cases where the average, median or maximum statistical power of primary studies that constitute the meta-analysis was larger than the power of meta-analysis.

		Average				**Median**				**Max**			
	k/μ_θ	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***
I^2= 40.43	***All***	1%	0%	1%		1%	2%	3%		9%	4%	4%	
	<10	3%	0%	3%		5%	0%	0%		29%	3%	3%	
	≥10	0%	0%	0%		0%	3%	4%		4%	4%	4%	
I^2= 62.61	***All***	12%	0%	0%		11%	3%	3%		62%	6%	4%	
	<10	48%	0%	0%		42%	3%	0%		94%	13%	6%	
	≥10	2%	0%	0%		3%	3%	4%		53%	4%	4%	
I^2= 79.46	***All***	64%	6%	1%		66%	6%	4%		89%	21%	7%	
	<10	94%	26%	3%		94%	26%	6%		100%	65%	19%	
	≥10	55%	0%	0%		58%	0%	4%		85%	8%	4%	
Post-hoc	***All***				11%				8%				21%
	<10				16%				16%				39%
	≥10				10%				5%				15%

Table 4. Percentage of cases where the average, median or maximum statistical power of primary studies that constitute the meta-analysis was larger than the power of meta-analysis.

		Average				**Median**				**Max**			
	k/μ_θ	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***	***0.2***	***0.5***	***0.8***	***Post-hoc***
I^2= 40.43	***All***	11%	3%	3%		10%	5%	5%		34%	9%	8%	
	<10	52%	13%	13%		45%	13%	10%		90%	29%	23%	
	≥10	0%	0%	0%		0%	3%	4%		18%	4%	4%	
I^2= 62.61	***All***	31%	4%	3%		28%	6%	6%		70%	15%	9%	
	<10	84%	19%	13%		84%	19%	13%		100%	55%	26%	
	≥10	16%	0%	0%		12%	3%	4%		61%	4%	4%	
I^2= 79.46	***All***	79%	15%	5%		76%	16%	7%		90%	35%	16%	
	<10	100%	61%	23%		97%	65%	19%		100%	84%	58%	
	≥10	73%	2%	0%		70%	2%	4%		87%	21%	4%	
Post-hoc	***All***				15%				15%				26%
	<10				26%				26%				52%
	≥10				12%				12%				19%

4.4 Discussion

The main aim of this study was to provide an overview of statistical power of random-effects meta-analyses on effectiveness of clinical psychological interventions. The statistical power of 141 meta-analyses was computed under various conditions of true effect and level of heterogeneity using different approaches. Additionally, we compared the statistical power of individual studies that contributed to these meta-analyses with the statistical power of the meta-analysis itself. Our results show that under certain scenarios, meta-analytic statistical testing is underpowered, even showing a lower statistical power than the average or maximum statistical power of included primary studies.

Our results revealed a clear impact of heterogeneity and the number of included primary studies on the statistical power of meta-analyses. While this pattern is expected, our findings show the steepness of this drop in statistical power. For instance, under low-heterogeneity, only 2% of meta-analyses had less than 80% power for detecting a true effect size of 0.5 using standard testing. However, this percentage increased to 11% and 31% under medium- and high-heterogeneity, respectively. The impact was even more pronounced for meta-analyses with less than 10 primary studies, with the percentage rising from 10% to 45% and 81%, respectively. Furthermore, in the case of Knapp-Hartung testing, as we should expect, the statistical power was lower compared to standard testing. For instance, almost half of the included meta-analyses with less than 10 primary studies exhibited a statistical power below 80% for detecting a true effect size of 0.8 under the high-heterogeneity scenario. This pattern aligns with expectations as the Knapp-Hartung test is designed to be more conservative. Simulation studies have shown that the type-1 error rate of the adjusted test is closer to the nominal level (Hartung and Knapp, 2001; InHout et al., 2014; Röver et al., 2015) compared to the standard test. Moreover, it has been advocated as conceptually more appropriate (van Aert and Jackson, 2019). Consequently, the results obtained from the Knapp-Hartung test can also be considered more accurate.

Basically, statistical power of the random-effects meta-analysis is a function of the true effect magnitude, which represents the expected value across studies under the random-effects model; the number of studies in the meta-analysis; the sample size of the primary studies; and the consistency across the results of included studies. Since the true effect is unknown and the number of included studies is determined by the available evidence at the time of conducting the meta-analysis, dealing with consistency becomes the most crucial task for a meta-analyst when employing a random-effects model. As our results have shown, the consistency across studies has a relevant impact on the statistical power of the summary effect test (and relatedly its precision). Furthermore, it also carries conceptual implications. Under a random-effects model it is assumed that the true effects of included studies are drawn from an underlying distribution from which the estimates of the primary studies are a random sample. In the presence of inconsistency across studies, conducting tests on the expected value across studies may be conceptually meaningless (Higgins et al., 2009), as some characteristics of the included studies are likely to vary systematically rather than randomly. In such situations, a natural step is to perform meta-regression analyses by extending the random-effects model to a mixed-effects model by including fixed covariates that may account for some of the inconsistency across effect sizes (Higgins et al., 2021). However, meta-regression tests also suffer from a of lack adequate statistical power when the number of included primary studies is low (Viechtbauer et al., 2015), requiring even a larger number of primary studies to achieve adequate statistical power (Cuijpers et al., 2021).

At this point, the inevitable question arises: How many studies are required to conduct a meta-analysis? In the context of discussing the statistical power of meta-analysis, Valentine et al. (2010) argued that even under circumstances where meta-analytic power may be low, any alternative synthesis method would be a worse choice than meta-analysis. Therefore, they conclude that the answer to that question is 2 studies. We could not agree more, since, as Valentine et al. (2010) argue, "given the need for some kind of synthesis, all the available alternatives are worse than meta-analysis, in that they are likely to be based on less defensible assumptions and on less transparent processes" (p. 239). However, it is important to bear in mind these limitations of random- and mixed-effects model, especially under circumstances such as low number of primary studies and inconsistency across studies.

On the other hand, when comparing the statistical power of individual studies to the power of meta-analyses, a similar pattern emerged. The percentage of cases where the average, median, or maximum statistical power of individual studies exceeded the power of the meta-analysis itself increased under the same circumstances where the power of the meta-analysis was lower. Furthermore, our study yielded similar results to those found by Jackson and Turner (2017) when calculating the statistical power of 1991 Cochrane reviews retrospectively, using the estimated values from the meta-analytic dataset as the true values. Ultimately, this type of comparison is more reflective of the trends in statistical power of the meta-analysis itself rather than the statistical power of the individual primary studies. After all, a primary study, and a meta-analysis answer essentially different question (regarding here a meta-analysis as the quantitative stage of a research synthesis project). Consequently, there is no scenario in which a primary study provides more information than a meta-analysis with respect to a research synthesis question. Nonetheless, there are scenarios, as discussed earlier, where certain results of a meta-analysis may be inaccurate, or lack sufficient power, such as the estimation of the summary effect and its statistical significance. Regardless, this comparison serves as a valuable illustration of how meta-analytic power behaves under specific circumstances.

References

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

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M., & Stanley, T. D. (2022). *Quantitative political science research is greatly underpowered* (No. 6). I4R

Borenstein, M. (2019) Heterogeneity in meta-analysis. In Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.), *The handbook of research synthesis and meta-analysis* (453-470). Russell Sage Foundation.

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

Carvalho, A. F., Solmi, M., Sanches, M., Machado, M. O., Stubbs, B., Ajnakina, O., ... & Herrmann, N. (2020). Evidence-based umbrella review of 162 peripheral biomarkers for major mental disorders. *Translational psychiatry*, *10*(1), 152.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological methods*, *8*(3), 243.

Cuijpers, P., Griffin, J., & Furukawa, T. (2021). The lack of statistical power of subgroup analyses in meta-analyses: A cautionary note. *Epidemiology and Psychiatric Sciences,* 30, E78.

de Vries, Y. A., Schoevers, R. A., Higgins, J. P., Munafò, M. R., & Bastiaansen, J. A. (2022). Statistical power in clinical trials of interventions for mood, anxiety, and psychotic disorders. *Psychological Medicine*, 1-8.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society open science*, *4*(2), 160254.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665-694.

Flint, J., Cuijpers, P., Horder, J., Koole, S. L., & Munafò, M. R. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression?. *Psychological medicine*, *45*(2), 439-446.

Guyatt, G. H., Mills, E. J., & Elbourne, D. (2008). In the era of systematic reviews, does the size of an individual trial still matter?. *PLoS medicine, 5*(1), e4.

Hartung, J. (1999). An alternative method for meta-analysis. Biometrical Journal: *Journal of Mathematical Methods in Biosciences, 41*(8), 901-916.

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in medicine, 20*(12), 1771-1782.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological methods*, *6*(3), 203.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558.

Higgins, J. P., López-López, J. A. & Aloe, A. M. (2021). Meta-regression. In Schmid, C. H., Stijnen, T., & White, I. (Eds.). *Handbook of meta-analysis.* CRC Press.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society, 172*(1), 137-159.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of selection and assessment, 8*(4), 275-292.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14, 1-12.

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research synthesis methods*, *8*(3), 290-302.

Jackson, D., Law, M., Rücker, G., & Schwarzer, G. (2017). The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in Medicine*, 36(25), 3923-3934.

Jia, P., Lin, L., Kwong, J. S., & Xu, C. (2021). Many meta-analyses of rare events in the Cochrane Database of Systematic Reviews were underpowered. *Journal of Clinical Epidemiology*, *131*, 113-122.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17), 2693-2710.

Konstantopoulos, S., & Hedges, L. V. (2019) Statistically Analyzing Effect Sizes: Fixed-and

Random-Effects Models. In Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.), *The handbook of research synthesis and meta-analysis* (245-280). Russell Sage Foundation.

Lopez-Nicolas, R., Lakens, D., López-López, J., Aparicio, M. R., Sandoval-Lentisco, A., López-Ibáñez, C., … Sánchez-Meca, J. (2022, November 25). Analytical reproducibility and data reusability of published meta-analyses on clinical psychological interventions. https://doi.org/10.31234/osf.io/gvqrn

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.

Niemeyer, H., Lorbeer, N., Mohr, J., Baer, E., & Knaevelsrud, C. (2022). Evidence-based individual psychotherapy for complex posttraumatic stress disorder and at-risk groups for complex traumatization: a meta-review. *Journal of affective disorders, 299*, 610-619.

Nuijten, M. B., Van Assen, M. A., Augusteijn, H. E., Crompvoets, E. A., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. *Journal of Intelligence*, *8*(4), 36.

Parish, A. J., Yuan, D. M., Raggi, J. R., Omotoso, O. O., West, J. R., & Ioannidis, J. P. (2021). An umbrella review of effect size, bias, and power across meta-analyses in emergency medicine. *Academic Emergency Medicine*, *28*(12), 1379-1388.

Röver, C., Knapp, G., & Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, *15*(1), 1-7.

Sánchez-Meca, J. & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48.

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2022). Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods*, *13*(1), 88-108.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*, *15*(3), e2000797.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. Research Synthesis Methods, 10(2), 180–194.

Turner, R. M., Bird, S. M., & Higgins, J. P. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PloS one*, *8*(3), e59202.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215-247.

van Aert, R. C., & Jackson, D. (2019). A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression. *Research Synthesis Methods, 10*(4), 515-527.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20*(3), 360–374.

Wiksten, A., Rücker, G., & Schwarzer, G. (2016). Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine, 35*(15), 2503-2515.

Chapter 5

General conclusions

Research synthesis projects play an indispensable role in the scientific process as they bring order to the vast array of scientific evidence, organizing individual pieces of evidence into a coherent body of knowledge on a specific topic. Given this prominent role, the results and conclusions of research synthesis projects carry greater relevance and impact compared to those of individual studies. Therefore, it is essential to keep an eye on the research practices and credibility of research synthesis projects. In this dissertation, we delve into various aspects of research practices and the credibility of research synthesis projects. The first study (Chapter 2) focused on assessing the prevalence of transparency and reproducibility-related reporting practices in research synthesis projects. The second study (Chapter 3) focused on reproducibility of meta-analytic results reported on these projects. Lastly, the third study (Chapter 4) explored the statistical power of meta-analytic synthesis when assuming a random-effects model. All three studies were carried out using a random sample of 100 published research synthesis projects on effectiveness of clinical psychological interventions.

In the first study, we comprehensively examined the entire process of a sample of research synthesis projects, from literature searching to synthesis methods. This study has been discussed in an earlier chapter. To sum up, we found major issues concerning completely reproducible search procedures report, specification of the exact method to compute effect sizes, or choice of weighting factors and estimators. Additionally, data availability was also examined and found a lack of availability of the statistics used to compute the effect sizes, as well as a lack of interoperability of available data.

In the second study, we delved deeper into the issue of data availability in research synthesis projects. Specifically, we assessed the reproducibility of the meta-analytic results reported in these works. Our findings indicated that current data sharing practices significantly hinder the reusability and retrieval of the data collected during a research synthesis project. This lack of accessible data emerged as one of the most significant threats to the reproducibility of meta-analytic results.

Nowadays, the benefits of data sharing are widely acknowledged. These benefits extend to all scientific endeavours, including research synthesis projects. However, we argue that proper data sharing is even more critical in the context of research synthesis. As discussed throughout this dissertation, one of the main aims of a research synthesis endeavour is to organize and bring order to a specific research topic. In this regard, the data collected through a well-conducted research synthesis project represents a comprehensive compilation of all the available evidence on a particular topic at a specific moment. Open data sharing of such collections holds tremendous value, as it encompasses a wealth of results, characteristics, study designs, and various other critical information from all the studies within a specific domain. Making this data openly accessible supports the intention to organise a specific area of research.

Moreover, the results from the third study have shown instances where random-effects meta-analytic averaging could be underpowered. As discussed in an earlier chapter, this does not invalidate the application of meta-analytical methods; however, it is essential to consider this aspect. Typically, meta-analytic averaging is extremely useful, providing more precise estimates and enabling more powerful statistical testing. Nevertheless, under specific circumstances, this type of synthesis may yield less informative results. While this highlights a limitation of meta-analytic averaging in some scenarios, it does not apply to research synthesis as a whole; rather, it is a specific outcome of the latter. In other words, research synthesis should not always aim to obtain an average effect as the primary result. Inconsistencies in results and limited evidence available on a particular topic are also relevant outcomes of synthesis work that shed light on the information available in a specific field, extending beyond the average effect.

To conclude, the specific conclusions drawn from this dissertation are detailed below:

- Several stages of published research synthesis projects lack sufficient reported information for reproducibility. For instance, completely reproducible search strategies were found to be limited in our reviewed papers, and there were also many cases where no information was provided on how dependency of coded primary data was dealt with. Furthermore, in several cases, crucial details of meta-analytic synthesis methods were insufficient. The computation formula for effect sizes or the estimation method for between-studies variance components was often not reported. (Chapter 2).

    - Completely reproducible search strategies are essential for both reproducing a research synthesis project from scratch and assessing the comprehensiveness of the synthesis. On the other hand, information on how dependency of coded primary results was dealt with, and the specific method chosen to compute effect sizes are crucial to be able to reproduce the unit of analysis of quantitative synthesis – the effect sizes.
- Meta-analytic shared data, when available, was typically limited to already computed effect sizes instead of primary data coded and used to compute them (Chapter 2 and 3).
    - Just sharing primary effect sizes might be problematic, as these outcomes result from an analytical process that has demonstrated numerous difficulties in reproducibility. These issues stem from factors like the aforementioned lack of information on how results were extracted and how effect sizes were computed.
- While there seems to have been an improvement in the availability of meta-analytic data over the years, frequently, the data accessible must still be extracted from document formats rather than being readily available as machine-readable files for reuse (Chapter 3).
    - Sharing data through document formats, such as tables or forest plots within the paper, hinders the data's reusability. This forces individuals who wish to reuse the data to engage in manual recoding, an inefficient and error-prone process. The availability of machine-readable data files, specifically designed for reuse, significantly enhances reusability. This aspect holds particular importance within the realm of research synthesis, where organized data is an inherent outcome of the project itself, as discussed earlier. The data extracted from a body of literature on a specific topic constitutes a valuable repository encompassing all available evidence pertinent to that particular subject.
- Overall, one of the biggest threats to the reproducibility of meta-analysis was related to data availability and current data sharing practices in meta-analysis. However, even when data retrieval was possible, some discrepancies were found in the results of some cases. We identified different issues that hindered our reproducibility attempts, such as reporting inconsistencies, lack of some data, or transcription errors (Chapter 3).

    - While it is worth mentioning that discrepancies in the results of these cases were generally minor and had minimal or no impact on the conclusions, it is important to address these avoidable situations. Some of the issues identified could be attributed to typographical errors. Fortunately, there are tools available that streamline the creation of what are known as reproducible manuscripts. These manuscripts incorporate analysis code, data, and result reporting into a single document, extracting and reporting results directly from the computational process's output, thereby mitigating the need for error-prone manual transcriptions.
- In specific scenarios, random-effects meta-analytic statistical testing demonstrates a lack of statistical power, sometimes even exhibiting lower power than the average or maximum statistical power of the included primary studies. These situations typically arise in the presence of high heterogeneity and a limited number of included primary studies (Chapter 4).
    - While these situations may affect the informativeness of the average effect, they do not diminish the intrinsic value of the synthesis effort. The characteristics inherent to the literature, which influence the inferential power regarding the average effect, stem from the current state of available research on a specific topic. Consequently, they represent an outcome of the synthesis itself.
- To summarize, we emphasize the role of research synthesis as an organizer of the research space. As discussed in Chapter 1, the demand for research synthesis arises within a context of explosive growth in the volume of available evidence, a trend that is even more pronounced in today's context. By highlighting this crucial dimension, the previously discussed issues become even more pertinent. As discussed in Chapter 2, completely reproducible systematic review methods, such as search strategies, are essential to rely on the body of evidence retrieved; As discussed in Chapter 3, openly sharing data files ready for reuse serves not only reproducibility concerns but also stands as a significant outcome of the organizational endeavour itself. Lastly, as discussed in Chapter 4, a single average result may not invariably be the most pertinent outcome of a research synthesis project.

Chapter 6

Resumen

Introducción general

El progreso del conocimiento científico se basa en la constante acumulación de conocimiento, en el trabajo desarrollado sobre las contribuciones previas. En un contexto de constante crecimiento de la cantidad de evidencia científica disponible, los enfoques de síntesis de la evidencia se tornan indispensables para alcanzar conclusiones sólidas y fiables. Sin embargo, a lo largo de la historia la forma de afrontar esta importante tarea ha variado. Desde revisiones narrativas, más subjetivas y no-sistemáticas, hasta enfoques sistemáticos, objetivos, transparentes y reproducibles, más acorde con las características de un proceso científico.

En la actualidad, las revisiones sistemáticas con meta-análisis han ganado reconocimiento como el *gold standard* en cuanto a síntesis de la evidencia se refiere. Harris M. Cooper y colaboradores a través de distintos trabajos (1982; 2017; 2019 delimitaron un proceso multietapa bien establecido, proporcionando un marco integral para la síntesis de la evidencia como un proceso científico. Este proceso se compone de las siguientes etapas: (1) Formulación del problema; (2) Selección de los estudios; (3) Extracción de información y evaluación crítica de la literatura; (4) Síntesis de la información; (5) Interpretación de los resultados; (6) Presentación y reporte.

Por otro lado, en la última década, distintas preocupaciones al respecto del proceso de producción científica han emergido en lo que ha sido considerado una crisis de credibilidad. Algunas contribuciones empíricas y conceptuales han señalado distintos problemas en este proceso. Dificultades a la hora de replicar distintos efectos observados en evidencia previa y la detección y alta prevalencia de distintas malas prácticas llevadas a cabo como *p-hacking*, *HARKing,* o sesgo de publicación han desembocado en una mayor atención en la investigación del propio proceso científico. En este contexto, se genera y desarrolla una disciplina científica centrada en la investigación de la propia investigación, conocida como meta-ciencia. Muchas han sido las contribuciones de esta disciplina en los últimos años. Sin embargo, la atención de este campo de investigación ha estado mayormente dirigida a la investigación primaria.

Objetivos

La presente tesis doctoral analiza empíricamente distintos aspectos relacionados con las buenas prácticas y la reproducibilidad de los trabajos de síntesis de la evidencia publicados. En primer lugar, se examinan las prácticas de reporte relacionadas con la transparencia y la reproducibilidad de las síntesis de investigación. En segundo lugar, investiga la reproducibilidad de los resultados reportados en síntesis cuantitativas (meta-análisis). Por último, evalúa la potencia estadística de los meta-análisis de efectos aleatorios. Estas evaluaciones se realizan sobre una muestra aleatoria de síntesis de investigación publicadas centradas en la efectividad de intervenciones de psicología clínica.

Una meta-revisión de la transparencia y practicas relacionadas con reproducibilidad en meta-análisis sobre intervenciones de psicología clínica

En este estudio, se evaluó empíricamente la prevalencia de las prácticas de reporte relacionadas con la transparencia y la reproducibilidad en meta-análisis sobre psicología clínica examinando una muestra aleatoria de 100 meta-análisis. Nuestro propósito fue identificar los puntos clave que podrían mejorarse, con el objetivo de proporcionar algunas recomendaciones para llevar a cabo meta-análisis reproducibles. Se realizó una meta-revisión de los meta-análisis de intervenciones psicológicas publicados entre 2000 y 2020. Se realizaron búsquedas en las bases de datos PubMed, PsycInfo y Web of Science. Se creó un formulario de codificación estructurado para evaluar los indicadores de transparencia basándose en estudios previos y en las directrices de meta-análisis existentes.

Reproducibilidad de meta-análisis sobre intervenciones de psicología clínica

En este estudio se evaluó la reproducibilidad de los resultados reportados de la muestra de meta-análisis publicados entre 2000-2020. De esta muestra, se seleccionaron 217 meta-análisis reportados en las publicaciones. En primer lugar, se intentó recuperar los datos originales recuperando un archivo de datos compartidos, recodificando los datos a partir de archivos de documentos o solicitándolos a los autores originales. En segundo lugar, mediante proceso multietapa, se intentó reproducir los principales resultados reportados de cada uno de los meta-análisis.

Potencia estadística de los meta-análisis sobre intervenciones de psicología clínica

Los estudios con poca potencia estadística son omnipresentes en psicología y disciplinas afines. El meta-análisis puede ayudar a paliar este problema, aumentando la potencia estadística al combinar los resultados de un conjunto de estudios primarios. Sin embargo, esto no es necesariamente cierto cuando utilizamos un modelo de efectos aleatorios, que es actualmente el enfoque predominante a la hora de realizar meta-análisis. En este estudio, examinamos la potencia estadística de una muestra de 141 meta-análisis de efectos aleatorios sobre la eficacia de las intervenciones psicológicas clínicas. Además, comparamos la potencia estimada de estos meta-análisis con la potencia de los estudios individuales que los componían. Para ello, utilizamos diferentes enfoques analíticos y un enfoque de Monte Carlo. La potencia estadística de los meta-análisis de efectos aleatorios se calculó bajo diferentes escenarios de tamaño del efecto verdadero y nivel de heterogeneidad.

Conclusiones generales

Los proyectos de síntesis de investigación desempeñan un papel indispensable en el proceso científico, ya que ordenan la vasta cantidad de evidencia científica, organizando las piezas individuales de evidencia en un cuerpo coherente de conocimiento sobre un tema específico. Dado este papel prominente, los resultados y conclusiones de los proyectos de síntesis de investigación tienen mayor relevancia e impacto en comparación con los estudios individuales. Por lo tanto, es esencial prestar atención a las prácticas de investigación y la credibilidad de estos proyectos. En esta tesis, se profundiza en varios aspectos de las prácticas de investigación y la credibilidad de los proyectos de síntesis.

El primer estudio (Capítulo 2) se centró en evaluar la prevalencia de prácticas de reporte relacionadas con la transparencia y la reproducibilidad en los proyectos de síntesis de investigación. El segundo estudio (Capítulo 3) se enfocó en la reproducibilidad de los resultados meta-analíticos reportados en estos proyectos. Por último, el tercer estudio (Capítulo 4) exploró la potencia estadística de la síntesis meta-analítica al asumir un modelo de efectos aleatorios. Los tres estudios se llevaron a cabo utilizando una muestra aleatoria de 100 proyectos de síntesis de investigación publicados sobre la efectividad de intervenciones psicológicas clínicas.

El primer estudio examinó de manera exhaustiva todo el proceso de una muestra de proyectos de síntesis de investigación, desde la búsqueda de literatura hasta los métodos de síntesis. Se encontraron problemas importantes en cuanto a la reproducibilidad de los procedimientos de búsqueda, la especificación del método exacto para calcular los tamaños de efecto o la elección de factores de ponderación y estimadores. Además, se examinó la disponibilidad de datos y se encontró una falta de disponibilidad de los estadísticos utilizados para computar los tamaños de efecto, así como una falta de interoperabilidad de los datos disponibles.

El segundo estudio profundizó en la cuestión de la disponibilidad de datos en los proyectos de síntesis de investigación. Específicamente, se evaluó la reproducibilidad de los resultados meta-analíticos reportados en estos trabajos. Se encontró que las prácticas actuales de compartimiento de datos dificultan significativamente la reutilización y recuperación de los datos recopilados durante un proyecto de síntesis de investigación. Esta falta de datos accesibles se convirtió en una de las amenazas más significativas para la reproducibilidad de los resultados meta-analíticos.

En la actualidad, se reconoce ampliamente los beneficios del intercambio de datos, y estos beneficios se extienden a todos los esfuerzos científicos, incluidos los proyectos de síntesis de investigación. Sin embargo, se argumenta que el intercambio adecuado de datos es aún más crítico en el contexto de la síntesis de investigación. Como se discutió a lo largo de esta tesis, uno de los principales objetivos de un trabajo de síntesis de investigación es organizar y ordenar un tema de investigación específico. En este sentido, los datos recopilados a través de un proyecto de síntesis de investigación bien realizado representan una compilación exhaustiva de toda la evidencia disponible sobre un tema específico en un momento dado. Compartir abiertamente estos datos tiene un valor altísimo, ya que abarca una gran cantidad de resultados, características, diseños de estudio y otra información crítica de todos los estudios dentro de un dominio específico. Hacer estos datos disponibles de forma abierta respalda la intención de organizar un área específica de investigación.

Además, los resultados del tercer estudio han mostrado casos en los que el promedio meta-analítico de efectos aleatorios podría tener poca potencia estadística.

Como se discutió en un capítulo anterior, esto no invalida la aplicación de métodos meta-analíticos; sin embargo, es esencial tener en cuenta este aspecto. Por lo general, el promedio meta-analítico es extremadamente útil, proporcionando estimaciones más precisas y permitiendo pruebas estadísticas más potentes. Sin embargo, bajo circunstancias específicas, este tipo de síntesis puede producir resultados menos informativos. Aunque esto resalta una limitación del promedio meta-analítico, no se aplica a la síntesis de investigación en su conjunto; más bien, es un resultado específico de esta. En otras palabras, la síntesis de investigación no siempre debe tener como objetivo principal obtener un efecto promedio como resultado. Las inconsistencias en los resultados y la evidencia limitada disponible sobre un tema en particular también son resultados relevantes del trabajo de síntesis que arrojan luz sobre la información disponible en un campo específico, que va más allá del efecto promedio.

En resumen, las conclusiones específicas extraídas de esta tesis son las siguientes:

- Varias etapas de proyectos de síntesis de investigación publicados carecen de información suficiente para ser reproducidos. Por ejemplo, se encontró que las estrategias de búsqueda completamente reproducibles eran limitadas en los proyectos revisados, y también hubo muchos casos en los que no se proporcionó información sobre cómo se manejó la dependencia de los datos primarios codificados. Además, en varios casos, los detalles cruciales de los métodos de síntesis meta-analítica fueron insuficientes.
- Los datos meta-analíticos compartidos, cuando estaban disponibles, eran típicamente limitados a tamaños de efecto ya calculados en lugar de datos primarios codificados y utilizados para calcularlos.
- Aunque parece existir una mejora en la disponibilidad de datos meta-analíticos a lo largo de los años, con frecuencia, los datos accesibles aún deben extraerse de archivos con formato de documento en lugar de estar listos para su reutilización como archivos legibles por máquinas.
- En general, una de las mayores amenazas para la reproducibilidad del meta-análisis estaba relacionada con la disponibilidad de datos y las prácticas actuales de intercambio de datos en el meta-análisis. Sin embargo, incluso cuando era posible recuperar los datos, se encontraron algunas discrepancias en los resultados.

- En situaciones específicas, las pruebas estadísticas meta-analíticas de efectos aleatorios muestran una falta de potencia estadística, a veces incluso exhibiendo una potencia menor que la potencia promedio o máximo de los estudios primarios incluidos.
- En conclusión, se enfatiza el papel de la síntesis de investigación como organizador del espacio de investigación. Los problemas discutidos se vuelven aún más pertinentes en este contexto, y se destaca la importancia de la transparencia en las prácticas de reporte, el intercambio de datos y la consideración de diferentes resultados en lugar de depender únicamente de un efecto promedio.

Appendices

Appendix 2A

Full search strategies and screening process summary

Supplementary Table 1. Full search strategy for each database

Database	Search strategy
PubMed	(meta-analy*[Title] OR “quantitative review“ OR “systematic review”[Title]) AND (psychotherap*[Title] OR "cognitive behavioral therapy"[Title] OR "behavior therapy"[Title] OR "cognitive behavioural therapy"[Title] OR "behaviour therapy"[Title] OR “CBT”[Title] OR "psychological treatments"[Title] OR "psychological interventions"[Title] OR "psychological treatment"[Title] OR "psychological intervention"[Title])
SCOPUS	TITLE(meta-analy* OR “quantitative review” OR “systematic review”) AND TITLE(psychotherap* OR "cognitive behavioral therapy" OR "behavior therapy" OR "cognitive behavioural therapy" OR "behaviour therapy" OR “CBT” OR "psychological treatments" OR "psychological interventions" OR "psychological treatment" OR

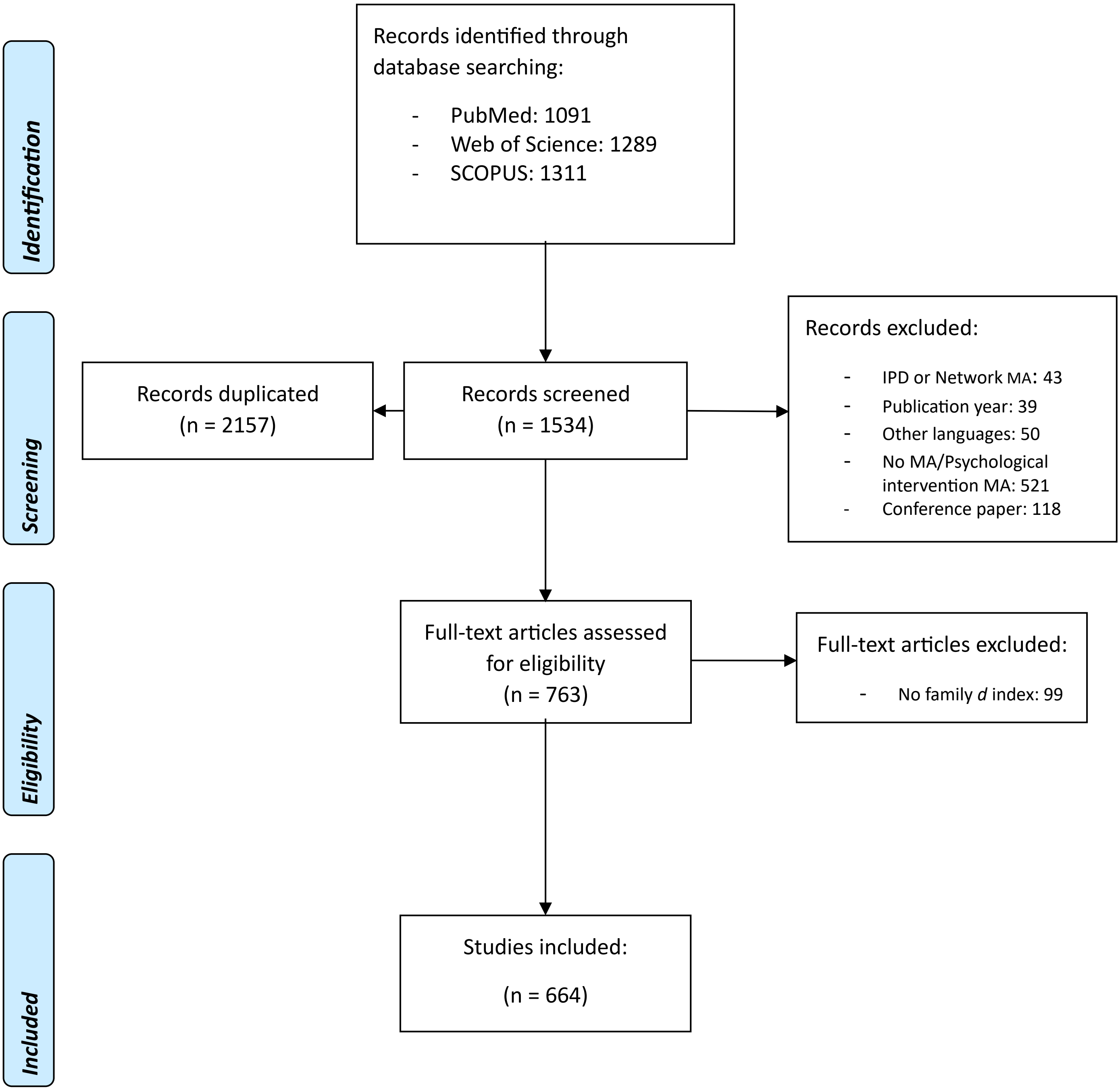
	"psychological intervention") AND PUBYEAR AFT 2000
Core collection of Web of Science	TI=(meta-analy* OR “quantitative review” OR “systematic review”) AND TI=(psychotherap* OR "cognitive behavioral therapy" OR "behavior therapy" OR "cognitive behavioural therapy" OR "behaviour therapy" OR “CBT” OR "psychological treatment" OR "psychological intervention" OR "psychological treatments" OR "psychological interventions"))

Search strategy development and previous exploratory searches

We developed the search strategy through iteration and discussion between authors of previous exploratory search outputs. The following terms were added throughout this process:

“systematic review”: We added this term in an attempt to capture some references that did not identify themselves in the title as a meta-analysis although they did carry out quantitative syntheses or meta-analyses.

“cognitive behavioral therapy” and “behavior therapy”: We added these terms to cover both British and American English spelling versions of “behaviour”.



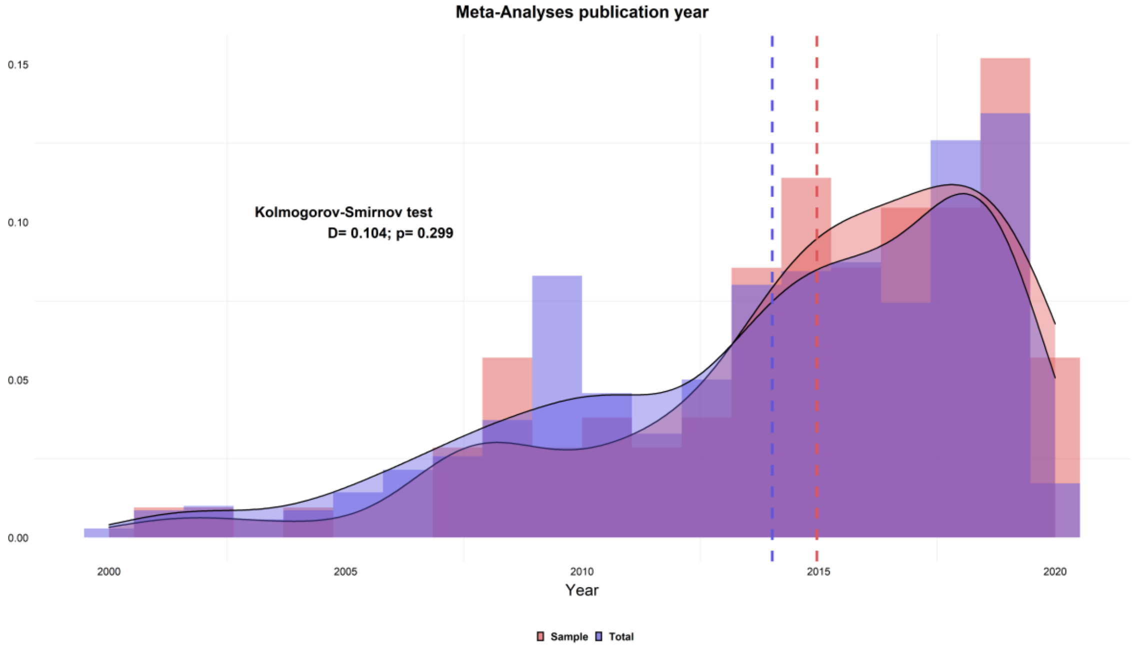
Supplementary Figure 1. Flow chart of the searching, screening, and selection process of the studies included in this study.

Considerations regarding the exclusion of 3 MAs from the selected random sample

From the random sample of 100 MAs selected, 3 studies had to be excluded for the following reasons:

- One of them was in a retracted state at the time of coding (Wang et al., 2019)
- Two could not be recovered in full text from any of the available sources (Magill et al., 2019; Proctor et al., 2018)

For this reason, three new meta-analyses were selected from the remaining ones not previously selected.



Supplementary Figure 2. Distribution of publication year for the included meta-analyses and for the selected random sample.

References

Magill, M., Ray, L., Kiluk, B., Hoadley, A., Bernstein, M., Tonigan, J. S., & Carroll, K. (2019). A meta-analysis of cognitive-behavioral therapy for alcohol or other drug use disorders: Treatment efficacy by contrast condition. *Journal of consulting and clinical psychology*, *87*(12), 1093-1105.

Proctor, B. J., Moghaddam, N., Vogt, W., & Das Nair, R. (2018). Telephone psychotherapy in multiple sclerosis: A systematic review and meta-analysis. *Rehabilitation psychology*, *63*(1), 16-28.

Wang, W., Zhou, Y., Chai, N., & Liu, D. (2019). Cognitive–behavioural therapy for personal recovery of patients with schizophrenia: A systematic review and meta-analysis. *General psychiatry*, *32*(4), e100040.

Appendix 2B

Kappa values for each item

Supplementary Table 2. Inter-coder agreement

Item	***kappa***	**N**
Item7	1	100
Item8	0.969	100
Item9	1	100
Item10	0.885	100
Item11	0.927	100
Item12	1	100
Item13	1	100
Item14	1	100
Item15	0.827	100
Item16	0.549	100
Item17	0.892	100
Item18	0.758	100
Item19	0.811	100
Item20	0.682	100
Item22	1	100
Item23	0.865	100
Item24	0.884	100
Item25	0.958	100
Item26	0.884	100
Item27	0.919	100
Item28	0.794	100
Item29	0.861	100
Item30	0.645	100
Item31	0.96	100
Item32	0.94	100
Item33	0.856	100
Item34	0.907	100
Item36	0.632	100
Item37	0.888	94

Item38	0.926	100
Item39	0.863	100
Item40	0.849	100
Item41	0.884	100
Item42	0.95	100
Item44	0.905	100
Item49	0.743	100
Item50	1	100
Item51	0.748	100
Item52	0.968	100
Item53	0.801	100
Item54	0.951	100
Item55	0.613	100
Item57	1	100
Item58	1	100
Item59	1	100
Item60	0.883	100
Item61	0.941	100
Item62	0.883	100

Appendix 2C

References of reviewed meta-analyses

Andersen, P., Toner, P., Bland, M., & McMillan, D. (2016). Effectiveness of Transdiagnostic Cognitive Behaviour Therapy for Anxiety and Depression in Adults: A Systematic Review and Meta-analysis. *Behavioural and Cognitive Psychotherapy*, *44*(6), 673-690. https://doi.org/10.1017/S1352465816000229

Arnberg, F. K., Linton, S. J., Hultcrantz, M., Heintz, E., & Jonsson, U. (2014). Internet-delivered psychological treatments for mood and anxiety disorders: A systematic review of their efficacy, safety, and cost-effectiveness. *PloS One*, *9*(5), e98118. https://doi.org/10.1371/journal.pone.0098118

Barak, A., Hen, L., Boniel-Nissim, M., & Shapira, N. (2008). *A comprehensive review and a meta-analysis of the effectiveness of Internet-based psychotherapeutic interventions*. Centre for Reviews and Dissemination (UK). https://www.ncbi.nlm.nih.gov/books/NBK76016/

Benish, S. G., Quintana, S., & Wampold, B. E. (2011). Culturally adapted psychotherapy and the legitimacy of myth: A direct-comparison meta-analysis. *Journal of Counseling Psychology*, *58*(3), 279-289. https://doi.org/10.1037/a0023626

Bird, V., Premkumar, P., Kendall, T., Whittington, C., Mitchell, J., & Kuipers, E. (2010). Early intervention services, cognitive-behavioural therapy and family intervention in early psychosis: Systematic review. *The British Journal of Psychiatry: The Journal of Mental Science*, *197*(5), 350-356. https://doi.org/10.1192/bjp.bp.109.074526

Birnie, K. A., Chambers, C. T., Taddio, A., McMurtry, C. M., Noel, M., Pillai Riddell, R., & Shah, V. (2015). Psychological Interventions for Vaccine Injections in Children and Adolescents: Systematic Review of Randomized and Quasi-Randomized Controlled Trials. *The Clinical Journal of Pain*, *31*, S72-S89. https://doi.org/10.1097/AJP.0000000000000265

Bortolotti, B., Menchetti, M., Bellini, F., Montaguti, M. B., & Berardi, D. (2008). Psychological interventions for major depression in primary care: A meta-analytic review of randomized controlled trials. *General Hospital Psychiatry*, *30*(4), 293-302. https://doi.org/10.1016/j.genhosppsych.2008.04.001

Boumparis, N., Karyotaki, E., Kleiboer, A., Hofmann, S. G., & Cuijpers, P. (2016). The effect of psychotherapeutic interventions on positive and negative affect in depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *202*, 153-162. https://doi.org/10.1016/j.jad.2016.05.019

Braun, S. R., Gregor, B., & Tran, U. S. (2013). Comparing bona fide psychotherapies of depression in adults with two meta-analytical approaches. *PloS One*, *8*(6), e68135. https://doi.org/10.1371/journal.pone.0068135

Briggs, S., Netuveli, G., Gould, N., Gkaravella, A., Gluckman, N. S., Kangogyere, P., Farr, R., Goldblatt, M. J., & Lindner, R. (2019). The effectiveness of psychoanalytic/psychodynamic psychotherapy for reducing suicide attempts and self-harm: Systematic review and meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *214*(6), 320-328. https://doi.org/10.1192/bjp.2019.33

Brown, L., Ospina, J. P., Celano, C. M., & Huffman, J. C. (2019). The Effects of Positive Psychological Interventions on Medical Patients' Anxiety: A Meta-analysis. *Psychosomatic Medicine*, *81*(7), 595-602. https://doi.org/10.1097/PSY.0000000000000722

Burns, A. M. N., Erickson, D. H., & Brenner, C. A. (2014). Cognitive-behavioral therapy for medication-resistant psychosis: A meta-analytic review. *Psychiatric Services (Washington, D.C.)*, *65*(7), 874-880. https://doi.org/10.1176/appi.ps.201300213

Captari, L. E., Hook, J. N., Hoyt, W., Davis, D. E., McElroy-Heltzel, S. E., & Worthington, E. L. (2018). Integrating clients' religion and spirituality within psychotherapy: A comprehensive meta-analysis: CAPTARI ET AL. *Journal of Clinical Psychology*, *74*(11), 1938-1951. https://doi.org/10.1002/jclp.22681

Carolan, S., Harris, P. R., & Cavanagh, K. (2017). Improving Employee Well-Being and Effectiveness: Systematic Review and Meta-Analysis of Web-Based Psychological Interventions Delivered in the Workplace. *Journal of Medical Internet Research*, *19*(7), e271. https://doi.org/10.2196/jmir.7583

Castro, A., Gili, M., Ricci-Cabello, I., Roca, M., Gilbody, S., Perez-Ara, M. Á., Seguí, A., & McMillan, D. (2020). Effectiveness and adherence of telephone-administered psychotherapy for depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *260*, 514-526. https://doi.org/10.1016/j.jad.2019.09.023

Cody, R. A., & Drysdale, K. (2013). *The effects of psychotherapy on reducing depression in residential aged care: A meta-analytic review*. Centre for Reviews and Dissemination (UK). https://www.ncbi.nlm.nih.gov/books/NBK127125/

Cristea, I. A., Gentili, C., Pietrini, P., & Cuijpers, P. (2017). Sponsorship bias in the comparative efficacy of psychotherapy and pharmacotherapy for adult depression: Meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *210*(1), 16-23. https://doi.org/10.1192/bjp.bp.115.179275

Cuijpers, P., Smit, F., & van Straten, A. (2007). Psychological treatments of subthreshold depression: A meta-analytic review. *Acta Psychiatrica Scandinavica*, *115*(6), 434-441. https://doi.org/10.1111/j.1600-0447.2007.00998.x

Cuijpers, Pim, Brännmark, J. G., & van Straten, A. (2008). Psychological treatment of postpartum depression: A meta-analysis. *Journal of Clinical Psychology*, *64*(1), 103-118. https://doi.org/10.1002/jclp.20432

Cuijpers, Pim, Geraedts, A. S., van Oppen, P., Andersson, G., Markowitz, J. C., & van Straten, A. (2011). Interpersonal psychotherapy for depression: A meta-analysis. *The American Journal of Psychiatry*, *168*(6), 581-592. https://doi.org/10.1176/appi.ajp.2010.10101411

Cuijpers, Pim, Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: A meta-analysis. *Journal of Affective Disorders*, *159*, 118-126. https://doi.org/10.1016/j.jad.2014.02.026

Cuijpers, Pim, Sijbrandij, M., Koole, S., Huibers, M., Berking, M., & Andersson, G. (2014). Psychological treatment of generalized anxiety disorder: A meta-analysis. *Clinical Psychology Review*, *34*(2), 130-140. https://doi.org/10.1016/j.cpr.2014.01.002

Cuijpers, Pim, van Straten, A., van Schaik, A., & Andersson, G. (2009). Psychological treatment of depression in primary care: A meta-analysis. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, *59*(559), e51-60. https://doi.org/10.3399/bjgp09X395139

Cuijpers, Pim, Weitz, E., Karyotaki, E., Garber, J., & Andersson, G. (2015). The effects of psychological treatment of maternal depression on children and parental functioning: A meta-analysis. *European Child & Adolescent Psychiatry*, *24*(2), 237-245. https://doi.org/10.1007/s00787-014-0660-6

Cusack, K., Jonas, D. E., Forneris, C. A., Wines, C., Sonis, J., Middleton, J. C., Feltner, C., Brownley, K. A., Olmsted, K. R., Greenblatt, A., Weil, A., & Gaynes, B. N. (2016). Psychological treatments for adults with posttraumatic stress disorder: A systematic review and meta-analysis. *Clinical Psychology Review*, *43*, 128-141. https://doi.org/10.1016/j.cpr.2015.10.003

Dahlenburg, S. C., Gleaves, D. H., & Hutchinson, A. D. (2019). Treatment outcome research of enhanced cognitive behaviour therapy for eating disorders: A systematic review with narrative and meta-analytic synthesis. *Eating Disorders*, *27*(5), 482-502. https://doi.org/10.1080/10640266.2018.1560240

Ghazanfarpour, M., Rajab Dizavandi, F., Kargarfard, L., Heidari, E., khadivzadeh, T., Saeidi, M., Abdolahian, S., & Safaei, M. (2018). Psychotherapy for Postpartum Depression in Iranian Women: A Systematic Review and Meta-Analysis. *International Journal of Pediatrics*, *6*(6). https://doi.org/10.22038/ijp.2018.30682.2690

Goldstein, E., McDonnell, C., Atchley, R., Dorado, K., Bedford, C., Brown, R. L., & Zgierska, A. E. (2019). The Impact of Psychological Interventions on Posttraumatic Stress Disorder and Pain Symptoms: A Systematic Review and Meta-Analysis. *The Clinical Journal of Pain*, *35*(8), 703-712. https://doi.org/10.1097/AJP.0000000000000730

Guarino, A., Polini, C., Forte, G., Favieri, F., Boncompagni, I., & Casagrande, M. (2020). The Effectiveness of Psychological Treatments in Women with Breast Cancer: A Systematic Review and Meta-Analysis. *Journal of Clinical Medicine*, *9*(1). https://doi.org/10.3390/jcm9010209

Heavens, D., Odgers, K., & Hodgekins, J. (2019). Cognitive behavioural therapy for anxiety in psychosis: A systematic review and meta-analysis. *Psychosis: Psychological, Social and Integrative Approaches*, *11*(3), 223-237. https://doi.org/10.1080/17522439.2019.1618380

Hershberger, A. R., Um, M., & Cyders, M. A. (2017). The relationship between the UPPS-P impulsive personality traits and substance use psychotherapy outcomes: A meta-analysis. *Drug and Alcohol Dependence*, *178*, 408-416. https://doi.org/10.1016/j.drugalcdep.2017.05.032

Hesser, H., Weise, C., Westin, V. Z., & Andersson, G. (2011). A systematic review and meta-analysis of randomized controlled trials of cognitive–behavioral therapy for tinnitus distress. *Clinical Psychology Review*, *31*(4), 545-553. https://doi.org/10.1016/j.cpr.2010.12.006

Hind, D., Cotter, J., Thake, A., Bradburn, M., Cooper, C., Isaac, C., & House, A. (2014). Cognitive behavioural therapy for the treatment of depression in people with multiple sclerosis: A systematic review and meta-analysis. *BMC Psychiatry*, *14*, 5. https://doi.org/10.1186/1471-244X-14-5

Huang, L., Zhao, Y., Qiang, C., & Fan, B. (2018). Is cognitive behavioral therapy a better choice for women with postnatal depression? A systematic review and meta-analysis. *PloS One*, *13*(10), e0205243. https://doi.org/10.1371/journal.pone.0205243

Huang, R., Yang, D., Lei, B., Yan, C., Tian, Y., Huang, X., & Lei, J. (2020). The short- and long-term effectiveness of mother-infant psychotherapy on postpartum depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *260*, 670-679. https://doi.org/10.1016/j.jad.2019.09.056

Ismail, K., Winkley, K., & Rabe-Hesketh, S. (2004). *Systematic review and meta-analysis of randomised controlled trials of psychological interventions to improve glycaemic control in patients with type 2 diabetes*. Centre for Reviews and Dissemination (UK). https://www.ncbi.nlm.nih.gov/books/NBK71115/

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

Kaddour, L., Kishita, N., & Schaller, A. (2018). A meta-analysis of low-intensity cognitive behavioral therapy-based interventions for dementia caregivers. *International Psychogeriatrics*, 1-16. https://doi.org/10.1017/S1041610218001436

Karyotaki, E., Smit, Y., de Beurs, D. P., Henningsen, K. H., Robays, J., Huibers, M. J. H., Weitz, E., & Cuijpers, P. (2016). THE LONG-TERM EFFICACY OF ACUTE-PHASE PSYCHOTHERAPY FOR DEPRESSION: A META-ANALYSIS OF RANDOMIZED TRIALS. *Depression and Anxiety*, *33*(5), 370-383. https://doi.org/10.1002/da.22491

Kazantzis, N., Whittington, C., Zelencich, L., Kyrios, M., Norton, P. J., & Hofmann, S. G. (2016). Quantity and Quality of Homework Compliance: A Meta-Analysis of Relations With Outcome in Cognitive Behavior Therapy. *Behavior Therapy*, *47*(5), 755-772. https://doi.org/10.1016/j.beth.2016.05.002

Klein, J. B., Jacobs, R. H., & Reinecke, M. A. (2007). Cognitive-behavioral therapy for adolescent depression: A meta-analytic investigation of changes in effect-size estimates. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*(11), 1403-1413. https://doi.org/10.1097/chi.0b013e3180592aaa

Kleinstäuber, M., Witthöft, M., & Hiller, W. (2011). Efficacy of short-term psychotherapy for multiple medically unexplained physical symptoms: A meta-analysis. *Clinical Psychology Review*, *31*(1), 146-160. https://doi.org/10.1016/j.cpr.2010.09.001

Koffel, E. A., Koffel, J. B., & Gehrman, P. R. (2015). A meta-analysis of group cognitive behavioral therapy for insomnia. *Sleep Medicine Reviews*, *19*, 6-16. https://doi.org/10.1016/j.smrv.2014.05.001

Kolubinski, D. C., Frings, D., Nikčević, A. V., Lawrence, J. A., & Spada, M. M. (2018). A systematic review and meta-analysis of CBT interventions based on the Fennell model of low self-esteem. *Psychiatry Research*, *267*, 296-305. https://doi.org/10.1016/j.psychres.2018.06.025

Koydemir, S., Sökmez, A. B., & Schütz, A. (2020). A Meta-Analysis of the Effectiveness of Randomized Controlled Positive Psychological Interventions on Subjective and Psychological Well-Being. *Applied Research in Quality of Life*. https://doi.org/10.1007/s11482-019-09788-z

Lee, H. J., Lee, J. H., Cho, E. Y., Kim, S. M., & Yoon, S. (2019). Efficacy of psychological treatment for headache disorder: A systematic review and meta-analysis. *The Journal of Headache and Pain*, *20*(1), 17. https://doi.org/10.1186/s10194-019-0965-4

Leichsenring, F., & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: Update of a meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *199*(1), 15-22. https://doi.org/10.1192/bjp.bp.110.082776

Lewis, C., Roberts, N. P., Simon, N., Bethell, A., & Bisson, J. I. (2019). Internet-delivered cognitive behavioural therapy for post-traumatic stress disorder: Systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, *140*(6), 508-521. https://doi.org/10.1111/acps.13079

Linardon, J., Fairburn, C. G., Fitzsimmons-Craft, E. E., Wilfley, D. E., & Brennan, L. (2017). The empirical status of the third-wave behaviour therapies for the treatment of eating disorders: A systematic review. *Clinical Psychology Review*, *58*, 125-140. https://doi.org/10.1016/j.cpr.2017.10.005

Linardon, J., Wade, T., de la Piedad Garcia, X., & Brennan, L. (2017). Psychotherapy for bulimia nervosa on symptoms of depression: A meta-analysis of randomized controlled trials. *International Journal of Eating Disorders*, *50*(10), 1124-1136. https://doi.org/10.1002/eat.22763

Malinauskas, R., & Malinauskiene, V. (2019). A meta-analysis of psychological interventions for Internet/smartphone addiction among adolescents. *Journal of Behavioral Addictions*, *8*(4), 613-624. https://doi.org/10.1556/2006.8.2019.72

Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., Bhullar, N., & Schutte, N. S. (2008). *Efficacy of cognitive behavioral therapy for chronic fatigue syndrome: A meta-analysis*. Centre for Reviews and Dissemination (UK). https://www.ncbi.nlm.nih.gov/books/NBK75484/

McDermut, W., Miller, I. W., & Brown, R. A. (2001). *The efficacy of group psychotherapy for depression: A meta-analysis and review of the empirical research*. Centre for Reviews and Dissemination (UK). https://www.ncbi.nlm.nih.gov/books/NBK68475/

Mitchell, L. J., Bisdounis, L., Ballesio, A., Omlin, X., & Kyle, S. D. (2019). The impact of cognitive behavioural therapy for insomnia on objective sleep parameters: A meta-analysis and systematic review. *Sleep Medicine Reviews*, *47*, 90-102. https://doi.org/10.1016/j.smrv.2019.06.002

Mohr, D. C., Vella, L., Hart, S., Heckman, T., & Simon, G. (2008). The Effect of Telephone-Administered Psychotherapy on Symptoms of Depression and Attrition: A Meta-Analysis. *Clinical psychology : a publication of the Division of Clinical Psychology of the American Psychological Association*, *15*(3), 243-253. https://doi.org/10.1111/j.1468-2850.2008.00134.x

Naeem, F., Khoury, B., Munshi, T., Ayub, M., Lecomte, T., Kingdon, D., & Farooq, S. (2016). Brief cognitive behavioral therapy for psychosis (CBTp) for schizophrenia: Literature review and meta-analysis. *International Journal of Cognitive Therapy*, *9*(1), 73-86. https://doi.org/10.1521/ijct_2016_09_04

Navarro-Bravo, B., Párraga-Martínez, I., López-Torres Hidalgo, J., Andrés-Pretel, F., & Rabanales-Sotos, J. (2015). Group cognitive-behavioral therapy for insomnia: A meta-analysis. [Terapia cognitivo-conductual grupal para el tratamiento del insomnio:

metaanálisis]. *Anales de Psicología*, *31*(1), 8-18. https://doi.org/10.6018/analesps.31.1.168641

Newman, E., Pfefferbaum, B., Kirlic, N., Tett, R., Nelson, S., & Liles, B. (2014). Meta-Analytic Review of Psychological Interventions for Children Survivors of Natural and Man-Made Disasters. *Current psychiatry reports*, *16*(9), 462. https://doi.org/10.1007/s11920-014-0462-z

Newton-Howes, G., & Wood, R. (2013). Cognitive behavioural therapy and the psychopathology of schizophrenia: Systematic review and meta-analysis. *Psychology and Psychotherapy*, *86*(2), 127-138. https://doi.org/10.1111/j.2044-8341.2011.02048.x

Ng, T. K., & Wong, D. F. K. (2018). The efficacy of cognitive behavioral therapy for Chinese people: A meta-analysis. *The Australian and New Zealand Journal of Psychiatry*, *52*(7), 620-637. https://doi.org/10.1177/0004867417741555

Okajima, I., & Inoue, Y. (2018). Efficacy of cognitive behavioral therapy for comorbid insomnia: A meta-analysis. *Sleep and Biological Rhythms*, *16*(1), 21-35. https://doi.org/10.1007/s41105-017-0124-8

Okuyama, T., Akechi, T., Mackenzie, L., & Furukawa, T. A. (2017). Psychotherapy for depression among advanced, incurable cancer patients: A systematic review and meta-analysis. *Cancer Treatment Reviews*, *56*, 16-27. https://doi.org/10.1016/j.ctrv.2017.03.012

Oldham, M., Kellett, S., Miles, E., & Sheeran, P. (2012). Interventions to increase attendance at psychotherapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *80*(5), 928-939. https://doi.org/10.1037/a0029630

Palpacuer, C., Gallet, L., Drapier, D., Reymann, J.-M., Falissard, B., & Naudet, F. (2017). Specific and non-specific effects of psychotherapeutic interventions for depression: Results from a meta-analysis of 84 studies. *Journal of Psychiatric Research*, *87*, 95-104. https://doi.org/10.1016/j.jpsychires.2016.12.015

Parker, G. B., Crawford, J., & Hadzi-Pavlovic, D. (2008). Quantified superiority of cognitive behaviour therapy to antidepressant drugs: A challenge to an earlier meta-analysis. *Acta Psychiatrica Scandinavica*, *118*(2), 91-97. https://doi.org/10.1111/j.1600-0447.2008.01196.x

Pearl, S. B., & Norton, P. J. (2017). Transdiagnostic versus diagnosis specific cognitive behavioural therapies for anxiety: A meta-analysis. *Journal of Anxiety Disorders*, *46*, 11-24. https://doi.org/10.1016/j.janxdis.2016.07.004

Perihan, C., Burke, M., Bowman-Perrott, L., Bicer, A., Gallup, J., Thompson, J., & Sallese, M. (2020). Effects of Cognitive Behavioral Therapy for Reducing Anxiety in Children with High Functioning ASD: A Systematic Review and Meta-Analysis. *Journal of Autism and Developmental Disorders*, *50*(6), 1958-1972. https://doi.org/10.1007/s10803-019-03949-7

Petrocelli, J. V. (2002). Effectiveness of Group Cognitive-Behavioral Therapy for General Symptomatology: A Meta-Analysis. *The Journal for Specialists in Group Work*, *27*(1), 92-115. https://doi.org/10.1177/0193392202027001008

Protogerou, C., Fleeman, N., Dwan, K., Richardson, M., Dundar, Y., & Hagger, M. S. (2015). Moderators of the effect of psychological interventions on depression and anxiety in cardiac surgery patients: A systematic review and meta-analysis. *Behaviour Research and Therapy*, *73*, 151-164. https://doi.org/10.1016/j.brat.2015.08.004

Reavell, J., Hopkinson, M., Clarkesmith, D., & Lane, D. A. (2018). Effectiveness of Cognitive Behavioral Therapy for Depression and Anxiety in Patients With Cardiovascular Disease: A Systematic Review and Meta-Analysis. *Psychosomatic Medicine*, *80*(8), 742-753. https://doi.org/10.1097/PSY.0000000000000626

Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clinical Psychology Review*, *32*(4), 329-342. https://doi.org/10.1016/j.cpr.2012.02.004

Rosa-Alcázar, A. I., Sánchez-Meca, J., Rosa-Alcázar, Á., Iniesta-Sepúlveda, M., Olivares-Rodríguez, J., & Parada-Navas, J. L. (2015). Psychological treatment of obsessive-compulsive disorder in children and adolescents: A meta-analysis. *The Spanish Journal of Psychology*, *18*, E20. https://doi.org/10.1017/sjp.2015.22

Ruiz, F. J. (2012). Acceptance and commitment therapy versus traditional cognitive behavioral therapy: A systematic review and meta-analysis of current empirical evidence. *International Journal of Psychology & Psychological Therapy*, *12*(3), 333-357.

Sánchez-Meca, J., Rosa-Alcázar, A. I., Iniesta-Sepúlveda, M., & Rosa-Alcázar, A. (2014). Differential efficacy of cognitive-behavioral therapy and pharmacological treatments for pediatric obsessive-compulsive disorder: A meta-analysis. *Journal of Anxiety Disorders*, *28*(1), 31-44. https://doi.org/10.1016/j.janxdis.2013.10.007

Schefft, C., Guhn, A., Brakemeier, E.-L., Sterzer, P., & Köhler, S. (2019). Efficacy of inpatient psychotherapy for major depressive disorder: A meta-analysis of controlled trials. *Acta Psychiatrica Scandinavica*, *139*(4), 322-335. https://doi.org/10.1111/acps.12995

Schmidt, H. M., Munder, T., Gerger, H., Frühauf, S., & Barth, J. (2014). Combination of psychological intervention and phosphodiesterase-5 inhibitors for erectile dysfunction: A narrative review and meta-analysis. *The Journal of Sexual Medicine*, *11*(6), 1376-1391. https://doi.org/10.1111/jsm.12520

Seyffert, M., Lagisetty, P., Landgraf, J., Chopra, V., Pfeiffer, P. N., Conte, M. L., & Rogers, M. A. M. (2016). Internet-Delivered Cognitive Behavioral Therapy to Treat Insomnia: A Systematic Review and Meta-Analysis. *PloS One*, *11*(2), e0149139. https://doi.org/10.1371/journal.pone.0149139

Sin, J., & Spain, D. (2017). Psychological interventions for trauma in individuals who have psychosis: A systematic review and meta-analysis. *Psychosis*, *9*(1), 67-81. https://doi.org/10.1080/17522439.2016.1167946

Smeets, K. C., Leeijen, A. A. M., van der Molen, M. J., Scheepers, F. E., Buitelaar, J. K., & Rommelse, N. N. J. (2015). Treatment moderators of cognitive behavior therapy to reduce aggressive behavior: A meta-analysis. *European Child & Adolescent Psychiatry*, *24*(3), 255-264. https://doi.org/10.1007/s00787-014-0592-1

Sockol, L. E. (2018). A systematic review and meta-analysis of interpersonal psychotherapy for perinatal women. *Journal of Affective Disorders*, *232*, 316-328. https://doi.org/10.1016/j.jad.2018.01.018

Spielmans, G. I., Pasek, L. F., & McFall, J. P. (2007). What are the active ingredients in cognitive and behavioral psychotherapy for anxious and depressed children? A meta-analytic review. *Clinical Psychology Review*, *27*(5), 642-654. https://doi.org/10.1016/j.cpr.2006.06.001

Straud, C. L., Siev, J., Messer, S., & Zalta, A. K. (2019). Examining military population and trauma type as moderators of treatment outcome for first-line psychotherapies for PTSD: A meta-analysis. *Journal of Anxiety Disorders*, *67*, 102133. https://doi.org/10.1016/j.janxdis.2019.102133

Sun, M., Rith-Najarian, L. R., Williamson, T. J., & Chorpita, B. F. (2019). Treatment Features Associated with Youth Cognitive Behavioral Therapy Follow-Up Effects for Internalizing Disorders: A Meta-Analysis. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *48*(sup1), S269-S283. https://doi.org/10.1080/15374416.2018.1443459

Tauber, N. M., O'Toole, M. S., Dinkel, A., Galica, J., Humphris, G., Lebel, S., Maheu, C., Ozakinci, G., Prins, J., Sharpe, L., Smith, A. "Ben", Thewes, B., Simard, S., & Zachariae, R. (2019). Effect of Psychological Intervention on Fear of Cancer Recurrence: A Systematic Review and Meta-Analysis. *Journal of Clinical Oncology*, *37*(31), 2899-2915. https://doi.org/10.1200/JCO.19.00572

Taylor, J. E., & Harvey, S. T. (2010). A meta-analysis of the effects of psychotherapy with adults sexually abused in childhood. *Clinical Psychology Review*, *30*(6), 749-767. https://doi.org/10.1016/j.cpr.2010.05.008

Thakral, M., Von Korff, M., McCurry, S. M., Morin, C. M., & Vitiello, M. V. (2020). Changes in dysfunctional beliefs about sleep after cognitive behavioral therapy for insomnia: A systematic literature review and meta-analysis. *Sleep Medicine Reviews*, *49*, 101230. https://doi.org/10.1016/j.smrv.2019.101230

Twomey, C., O'Reilly, G., & Byrne, M. (2015). Effectiveness of cognitive behavioural therapy for anxiety and depression in primary care: A meta-analysis. *Family Practice*, *32*(1), 3-15. https://doi.org/10.1093/fampra/cmu060

Twomey, C., O'Reilly, G., & Meyer, B. (2017). Effectiveness of an individually-tailored computerised CBT programme (Deprexis) for depression: A meta-analysis. *Psychiatry Research*, *256*, 371-377. https://doi.org/10.1016/j.psychres.2017.06.081

Uchendu, C., & Blake, H. (2017). Effectiveness of cognitive-behavioural therapy on glycaemic control and psychological outcomes in adults with diabetes mellitus: A systematic review and meta-analysis of randomized controlled trials. *Diabetic Medicine: A Journal of the British Diabetic Association*, *34*(3), 328-339. https://doi.org/10.1111/dme.13195

van der Zweerde, T., Bisdounis, L., Kyle, S. D., Lancee, J., & van Straten, A. (2019). Cognitive behavioral therapy for insomnia: A meta-analysis of long-term effects in controlled studies. *Sleep Medicine Reviews*, *48*, 101208. https://doi.org/10.1016/j.smrv.2019.08.002

van Dis, E. A. M., van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L. H., van den Heuvel, R. M., Cuijpers, P., & Engelhard, I. M. (2020). Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *77*(3), 265-273. https://doi.org/10.1001/jamapsychiatry.2019.3986

van Hees, M. L. J. M., Rotter, T., Ellermann, T., & Evers, S. M. A. A. (2013). The effectiveness of individual interpersonal psychotherapy as a treatment for major depressive disorder in adult outpatients: A systematic review. *BMC Psychiatry*, *13*(1), 22. https://doi.org/10.1186/1471-244X-13-22

Vázquez, F. L., Hermida, E., Díaz, O., Torres, Á., Otero, P., & Blanco, V. (2014). Intervenciones psicológicas para cuidadores con síntomas depresivos: Revisión sistemática y metanálisis. *Revista Latinoamericana de Psicología*, *46*(3), 178-188. https://doi.org/10.1016/S0120-0534(14)70021-4

Whiteside, S. P. H., Sim, L. A., Morrow, A. S., Farah, W. H., Hilliker, D. R., Murad, M. H., & Wang, Z. (2020). A Meta-analysis to Guide the Enhancement of CBT for Childhood Anxiety: Exposure Over Anxiety Management. *Clinical Child and Family Psychology Review*, *23*(1), 102-121. https://doi.org/10.1007/s10567-019-00303-2

Windsor, L. C., Jemal, A., & Alessi, E. J. (2015). Cognitive behavioral therapy: A meta-analysis of race and substance use outcomes. *Cultural Diversity & Ethnic Minority Psychology*, *21*(2), 300-313. https://doi.org/10.1037/a0037929

Woll, C. F. J., & Schönbrodt, F. D. (2020). A series of meta-analytic tests of the efficacy of long-term psychoanalytic psychotherapy. *European Psychologist*, *25*(1), 51-72. https://doi.org/10.1027/1016-9040/a000385

Xiang, X., Wu, S., Zuverink, A., Tomasino, K. N., An, R., & Himle, J. A. (2020). Internet-delivered cognitive behavioral therapies for late-life depressive symptoms: A systematic review and meta-analysis. *Aging & Mental Health*, *24*(8), 1196-1206. https://doi.org/10.1080/13607863.2019.1590309

Yang, L., Zhou, X., Zhou, C., Zhang, Y., Pu, J., Liu, L., Gong, X., & Xie, P. (2017). Efficacy and Acceptability of Cognitive Behavioral Therapy for Depression in Children: A Systematic Review and Meta-analysis. *Academic Pediatrics*, *17*(1), 9-16. https://doi.org/10.1016/j.acap.2016.08.002

Ye, Y., Zhang, Y., Chen, J., Liu, J., Li, X., Liu, Y., Lang, Y., Lin, L., Yang, X.-J., & Jiang, X.-J. (2015). Internet-Based Cognitive Behavioral Therapy for Insomnia (ICBT-i) Improves Comorbid Anxiety and Depression—A Meta-Analysis of Randomized Controlled Trials. *PLOS ONE*, *10*(11), e0142258. https://doi.org/10.1371/journal.pone.0142258

Zhang, Q., Jiang, S., Young, L., & Li, F. (2019). The Effectiveness of Group-Based Physiotherapy-Led Behavioral Psychological Interventions on Adults With Chronic Low Back Pain: A Systematic Review and Meta-Analysis. *American Journal of Physical Medicine & Rehabilitation*, *98*(3), 215-225. https://doi.org/10.1097/PHM.0000000000001053

Appendix 3A

Inter-coder agreement

Out of the 25 papers selected to carry out independent double-coding, 21 had primary data available for one or more meta-analyses selected under the criteria explained in the main manuscript. Some disagreements or coding errors in the primary data were found in 8 (38%) cases. The intraclass correlation coefficient between the values coded by each coder was computed for each of those 21 datasets. The ICC values varied between 0.988 and 1 (mean = 0.999). Full results of the inter-coder agreement are presented in Table S3.

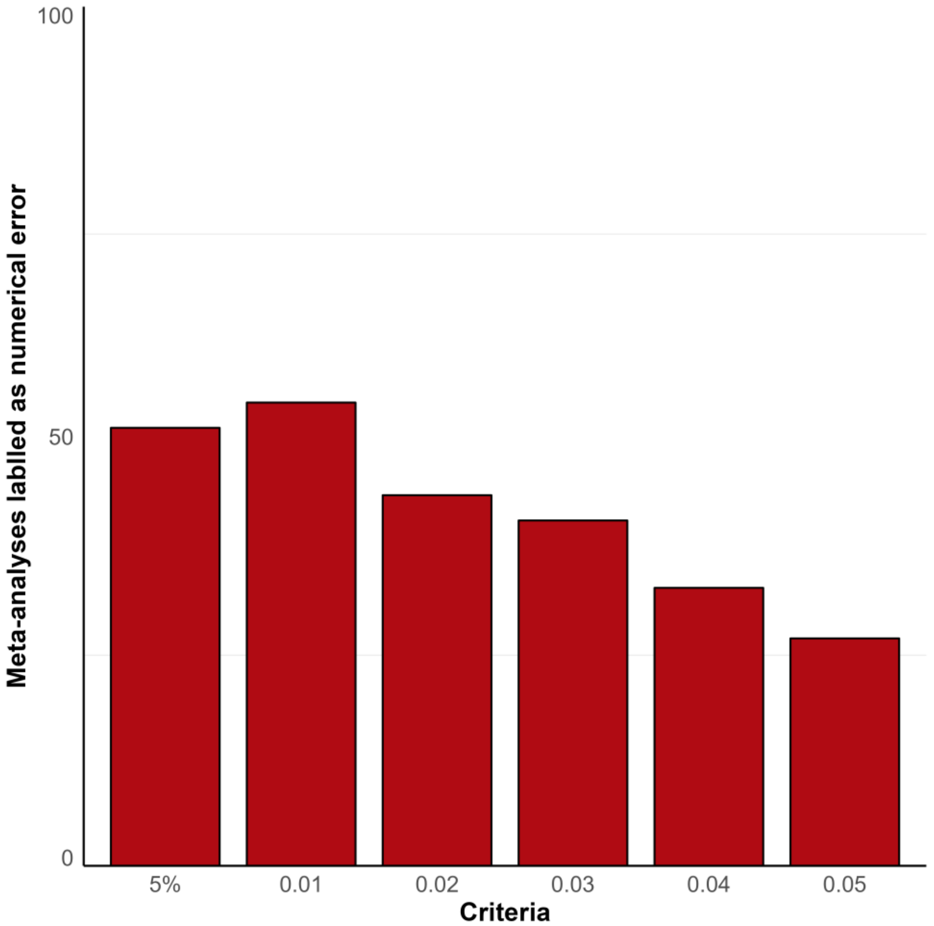
Supplementary Table 3. Full results of the inter-coder agreement

Paper	ICC	Full agreement
1	.988	FALSE
2	1	TRUE
3	1	FALSE
4	1	FALSE
5	1	TRUE
6	1	TRUE
7	.997	FALSE
8	1	TRUE
9	1	FALSE
10	1	TRUE
11	1	TRUE
12	.999	FALSE
13	1	TRUE
14	1	TRUE
15	1	TRUE
16	1	TRUE
17	1	TRUE
18	1	FALSE
19	.977	FALSE
20	1	TRUE
21	1	TRUE

Appendix 3B

Sensitivity analysis using other possible criteria

In the first stage, a cut-off point of 5% discrepancy between reported and reproduced results was set to screen meta-analyses. In Figure S3 this criterion is compared with other possibilities in absolute value. Figure S3 shows that the criterion used is one of the most liberal of those compared. Due to our design, the stage where the criterion was applied is only an initial screening, so the meta-analyses labelled by this criterion were reviewed at later stages in a qualitative way. Therefore, a more liberal criterion (minimizing false negatives, but increasing false positives) is more suitable. This criterion has the limitations of any relative index, but due to the mix of metrics included, it is also considered more appropriate.



Supplementary Figure 3. Barplot displaying the number of meta-analyses labelled as numerical error under different criteria. The value used in the original design is 5%, the other bars correspond to alternative criteria of the difference in absolute value between reported and reproduced result.

Appendix 3C

Data availability over years

The publication year range of the included meta-analyses in our study is quite extensive. The initial pool, from which we randomly selected a sample, comprised meta-analytic reports published between 2000 and 2020, encompassing a span of two decades. For our analysis, we focused on a random sample of 100 of these meta-analytic reports. From the 100 included reports, 217 independent meta-analyses were selected following the criteria explained in the 'Identification and selection of articles and meta-analyses' section of the main paper. These meta-analyses were published between 2001 and 2020 (mean = 2015.04; sd = 4.05; median = 2016; interquartile range = 2012-2016). As an unrestricted random sample, the publication year distribution is clearly left-skewed, with 92.2% of the included meta-analyses published between 2010 and 2020 (see Figure 2 of the main paper for the full distribution).

As shown in the main text (see Figure 3a), there appears to be an improvement in the rate of data availability over the years. The overall data availability rate for the full sample was found to be 67%. However, when examining meta-analyses published within specific time periods, the rates varied. For meta-analyses published between 2000 and 2010, the data availability rate was 41%. For those published between 2011 and 2015, the rate increased to 59%. Notably, meta-analyses published between 2016 and 2020 exhibited the highest data availability rate of 80%. These findings suggest a positive trend of improved data availability in more recent years.

The association between data availability and publication year was 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. The analyses of the main manuscripts were mostly carried out at meta-analysis level. However, our dataset has a nested structure with meta-analyses nested within papers, which could compromise the assumption of independence of the regression model. Hence, we fitted the regression model at both meta-analysis and paper levels. Since we selected more than one meta-analysis from some of the papers, in 7 cases primary data only could be retrieved for certain of the meta-analyses in that paper, but not for all the selected meta-analyses in that paper. To avoid misclassifications, these cases were excluded from the paper-level analyses. Furthermore, as we work on an unrestricted random sample, the publication year distribution of the sample is clearly left-skewed. Therefore, the information provided by the data at the bottom range of the predictor is limited. For this reason, the analyses were complemented by fitting logistic regression models on a subset of the data excluding the papers published before 2010. In summary, we fitted four binary logistic regression model at both meta-analysis and paper levels and using the full dataset or a subset of meta-analyses published after 2010. Despite multiple contrasts performed, we did not introduce any corrections for multiple comparisons due to the exploratory nature of our analyses. Table S4 summarises the results of the models.

Based on the results of the four models, there seems to be an association between the publication year and the possibility of retrieving primary data from a meta-analysis. We found that all 4 ORs computed were > 1, indicating a higher probability of retrieving data the more recent the publication year of the meta-analysis. Specifically, the odds increased from 11.85% to 34.46% per year. Additionally, in only one of the cases (paper level model without excluding cases published before 2010) did the 95%CI of the OR include a value < 1.

Supplementary Table 4. Binary logistic regression models results

Level	Before 10s exclusion	Slope	OR	OR LL	OR UL	*p*	Percentage change
Meta-analysis	No	0.155	1.168	1.086	1.262	0.000	16.769
Paper	No	0.112	1.119	0.994	1.260	0.059	11.852
Meta-analysis	Yes	0.296	1.345	1.212	1.503	0.000	34.464
Paper	Yes	0.250	1.284	1.026	1.633	0.032	28.351

Appendix 3D

Qualitative check results

Supplementary Table 5. Reasons found during qualitative assessment.

Case	Reason
1	No clear reason was found. Labelled as numerical error due to discrepancy in upper confidence limit. This difference (0.029) was considered not relevant.
2	Inverted signs of results. Can be explained by authors choosing to report absolute values in the main text.
3	Inverted signs of results. Can be explained by authors choosing to report absolute values in the main text.
4	No clear reason was found. Labelled as numerical error due to discrepancy in upper confidence limit. This difference (0.017) was considered not relevant.
5	Reproduced values rounded to two decimal places match with the original results (reported rounded to two decimal places).
6	Original results extracted from a subgroup analysis. Results match using a pooled between-studies variance estimation instead of a separate one.
7	No clear reason was found. Labelled as numerical error due to discrepancy in lower confidence limit. This difference (0.023) was considered not relevant.
8	No clear reason was found. Labelled as numerical error due to discrepancy in upper confidence limit. This difference (0.016) was considered not relevant.
9	Reproduced values rounded to one decimal place match with the original results (reported rounded to one decimal place).
10	Inverted signs of results. Can be explained by authors choosing to report absolute values in the main text.
11	Inverted signs of results. Can be explained by authors choosing to report absolute values in the main text.
12	No clear reason was found. Labelled as numerical error due to discrepancy in lower confidence limit. This difference (0.026) was considered not relevant.
13	No clear reason was found. Labelled as numerical error due to discrepancy in upper confidence limit. This difference (0.026) was considered not relevant.
14	Reproduced values rounded to two decimal places match with the original results (reported rounded to two decimal places).
15	The results match if the sign of the upper confidence interval is reversed. It was considered a minor reporting error.