UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Magistrale in
Psicologia di Comunità, della Promozione del Benessere e del Cambiamento Sociale

Tesi di Laurea Magistrale

**A new method to explore inferential risks
associated with each study in a meta-analysis: An
approach based on Design Analysis**

**Un nuovo metodo per esplorare i rischi inferenziali di ogni
studio di una meta analisi: Un approccio basato sulla Design
Analysis**

Relatore:
Prof. Gianmarco Altoè

Laureanda: Francesca Giorgi
Matricola: 1228012

Anno Accademico 2020/2021

Ai miei nonni

Ringraziamenti

Un sentito ringraziamento al Prof. Altoè, per la disponibilità e il supporto nella stesura di questa tesi. Grazie per la passione e la motivazione che mi ha trasmesso in questi mesi, i suoi insegnamenti saranno sempre di ispirazione per il mio lavoro.

A Chiara, Sara, Marina, Gianluca, Ambra e Denise, per aver condiviso con me gioie, frustrazioni, delusioni e successi. Avete reso questo travagliato, ma meraviglioso viaggio bello da morire. Siete il regalo più bello che questa esperienza mi abbia donato.

Alle amiche della Valle, Annalisa, Martina, Sharon e Giada per essermi state vicino in questo anno difficile, per avermi dato la forza e la motivazione di non arrendermi.

Un grazie speciale alla mia famiglia, per essere stati la spalla su cui piangere e il viso a cui sorridere.

Grazie al mio papà per avermi sostenuto sempre, per esserti preso cura di me anche quando eravamo distanti e per avermi lasciata libera di scegliere la mia strada, appoggiando le mie decisioni e offrendomi il tuo supporto.

Grazie a mio fratello, per esserci stato quando nessun altro c'era, per essere stato mio amico, sostegno e confidente. Grazie per avermi trasmesso il coraggio di essere chi sono, senza vergogna e senza paura del giudizio degli altri, perchè tutti meritano e hanno il diritto di essere amati per quello che sono.

Grazie alla mia mamma, per essere stata il mio punto di riferimento, il modello a cui ispirarmi e il porto sicuro in cui tornare. Grazie per aver mantenuto la promessa che mi feci 5 anni fa, "qualsiasi cosa succeda, non preoccuparti, ci siamo noi, tu vai avanti che noi ci occupiamo del resto". Sono sicura che non sarei qui se quelle parole fossero venute meno, per questo a voi va il mio grazie più grande.

Table of Contents

Riassunto	7
1 Psychology's Credibility Crisis	11
1.1 Psychology's credibility crisis	11
1.1.1 Questionable Research Practices	13
1.1.2 Questionable Measurement Practices	16
1.1.3 Low Statistical power	18
1.2 Replicability and Reproducibility	21
1.3 Pre-registrations and Open Science Framework	25
1.3.1 Aims	27
2 Design Analysis	29
2.1 The design analysis	29
2.1.1 Type M error, Type S error and Statistical Power	31
2.1.2 Plausible effects size	33
2.2 Prospective and retrospective design analysis	35
2.2.1 Prospective design analysis	35
2.2.2 Retrospective design analysis	37
3 The Meta-Analysis	39
3.1 Meta-analysis	39
3.2 The Benefits Of Conducting a Meta-Analysis	40
3.3 How To Conduct a Meta-Analysis	42
3.3.1 Calculating An Effect Size For Each Study	43
3.3.2 Summarizing Effects: Fixed-Effect Versus Random-Effects Models	44
3.3.3 Transformation of effect-sizes: Fisher's z transformation	47
3.3.4 Heterogeneity Assessment	47
3.3.5 Publication Bias Evaluation	49
4 drmeta	53
4.1 Introduction	54
4.2 <i>drmeta</i> Function: How It Works	55
4.2.1 drmeta Application: Input and Output	57

5	A Case Study	63
5.1	The Original Study	63
5.2	Exploring The Inferential Risks Associated With Each Study In The Meta-Analysis: Application On A Published Study	65
5.3	Further Potentials Of The drmeta Function	70
6	Conclusions	73
	References	77
	A drmeta Function	
	B How drmeta function works	
	C The Case study	

Riassunto

Negli ultimi dieci anni, la ricerca scientifica, e in particolare la ricerca in psicologia, sta vivendo quella che viene chiamata crisi di credibilità dei risultati. Quello che si è verificato è stata una difficoltà da parte dei ricercatori di trovare gli stessi risultati di studi presenti in letteratura replicandone la procedura: spesso gli effetti non risultavano così forti come negli studi originari e qualche volta nello studio di replica non veniva rilevato alcun effetto. Questa crisi di riproducibilità e replicabilità dei risultati, da una parte, ha aumentato la consapevolezza dei ricercatori rispetto ai problemi presenti nella letteratura e dall'altra, ha avuto un importante effetto collaterale nel promuovere lo sviluppo di nuove pratiche che garantissero rigore, trasparenza e riproducibilità.

Nel presente lavoro, l'obiettivo è di proporre un nuovo metodo per esplorare i rischi inferenziali di ogni studio di una meta-analisi. In particolare, questo metodo è basato sulla Design Analysis, un approccio di analisi della potenza statistica ideato da Gelman & Carlin (2014), che permette di analizzare due errori comunemente non considerati: l'errore di Tipo M (Magnitude) e l'errore di Tipo S (Sign), riguardanti la grandezza e la direzione degli effetti. Abbiamo scelto l'approccio della Design Analysis, perchè permette di porre maggior enfasi sulla stima della dimensione dell'effetto e perchè può essere uno valido strumento per aumentare la coscienziosità del ricercatore rispetto allo studio che andrà ad implementare.

Nel primo capitolo viene spiegato cosa intendiamo con crisi di credibilità in psicologia. Vengono quindi analizzate le origini e i principali fattori che sono ritenuti causa

della crisi e le possibili soluzioni avanzate per promuovere un cambiamento. Il secondo capitolo presenta la Design Analysis. In particolare, vengono analizzati gli errori di Tipo M e di Tipo S e il loro ruolo nella crisi di credibilità. Infatti, studi che presentano alte probabilità di commettere questi errori, potrebbero fornire stime di effetti la cui grandezza è esagerata e/o nella direzione sbagliata. Inoltre, vengono presentate la *prospective design analysis*, che permette di esplorare i rischi inferenziali prima di condurre uno studio, e la *retrospective design analysis*, che permette di esplorare i rischi inferenziali di uno studio che è già stato condotto. Nel terzo capitolo viene spiegato cos'è, quali sono i benefici e come si conduce uno studio meta-analitico, con particolare enfasi sull'analisi statistica. La meta-analisi rappresenta infatti uno strumento importante a disposizione dei ricercatori per aumentare la fiducia rispetto alla dimensione dell'effetto studiato e alla precisione; inoltre può essere usata per stimare la dimensione dell'effetto plausibile nelle fasi iniziali di uno studio. Nel quarto capitolo viene presentato il nostro metodo per calcolare i rischi inferenziali di ogni studio di una meta-analisi. In particolare, tramite la nostra funzione *drmeta* è possibile calcolare in R la potenza, l'errore di tipo M e l'errore di tipo S attraverso cinque diversi metodi: Standard, Standard Corrected, All Studies, All Studies Corrected e Personalizzato. Viene quindi spiegato come funziona *drmeta* e, attraverso un semplice esempio, viene mostrato come interpretare i risultati. Si propone un case study per mostrare le potenzialità della nostra funzione nell'analisi di uno studio meta-analitico. L'ultimo capitolo riassume i risultati del presente lavoro, enfatizzando il ruolo della design analysis nell'analisi dei risultati pubblicati in letteratura.

Il presente lavoro è stato scritto in RMarkdown.

Summary

In the last ten years, scientific research, and specifically scientific research in psychology, has experienced an unprecedented “credibility’s crisis” of results. This means that researchers couldn’t find the same results as the original studies when conducting replication studies. The results showed that effects size were often not as strong as in the original studies and sometimes no effect was found. However, an important side-effect of the replicability crisis is that it increased the awareness of the problematic issues in the published literature and it promoted the development of new practices which would guarantee rigour, transparency and reproducibility.

In the current work, the aim is to propose a new method to explore the inferential risks associated with each study in a meta-analysis. Specifically, this method is based on Design Analysis, a power analysis approach developed by Gelman & Carlin (2014), which allows to analyse two other type of errors that are not commonly considered: the Type M (Magnitude) error and the Type S (Sign) error, concerning the magnitude and direction of the effects. We chose the Design Analysis approach because it allows to put more emphasis on the estimate of the effect size and it can be a valid tool available to researchers to make more conscious and informed decisions.

The first chapter explains what we mean by psychology credibility’s crisis. The focus is on the origins, the main factors that are considered to be the cause of the crisis and the solution proposed for a change of direction. The second chapter presents the Design Analysis. Specifically, Type M and Type S errors and their role in the credibility crisis are analysed. In fact, studies with a high probability of committing these errors could provide effect size estimates whose magnitude is exaggerated and/or in the wrong direction. In addition, *prospective design analysis*, which allows to explore inferential risks before conducting a study, and *retrospective design analysis*, which allows to explore the inferential risks of a study that has already been conducted, are presented. The third chapter explains what is a meta-analysis, the benefits of

conducting a meta-analytic study, and how to conduct it, with particular emphasis on the statistical analysis. In fact, meta-analysis represents an important tool for researchers to increase confidence in the effect size and on the precision (i.e., confidence interval) of the study and it can be used to estimate plausible effect sizes in the early stages of a study. In the fourth chapter, we present our method to calculate the inferential risks associated with each study in a meta-analysis. Specifically, using our *drmeta* function, it is possible to calculate power, M-type error and S-type error in R through five different methods: Standard, Standard Corrected, All Studies, All Studies Corrected and Personalized. We explain how *drmeta* works and, through a simple example, we show how the results should be interpreted. Then, a case study illustrates the potential associated with our function. The last chapter summarizes the results of this work, emphasizing the role of design analysis in analyzing results published in the literature.

The current work was written in Rmarkdown.

Chapter 1

What we mean by Psychology's Credibility crisis

In this chapter, we will explain what we mean by psychology's credibility crisis. At the beginning, we will present different factors hypothesized to be responsible for the low replication rate (e.g. Questionable Research Practices and Questionable Measurement Practices) with the focus especially on statistical power. Next, we will discuss the importance of reproducibility and replicability in the scientific progress with emphasis on the role of replication and the interpretation of possible outcomes. At the end, we will illustrate some practices that have been proposed as some of the possible solution to the credibility crisis. This chapter will conclude with a brief presentation of the aim of this thesis.

1.1 Psychology's credibility crisis

In the last twelve years, social science, as well as biology, medicine and other statistic-dependent fields, has experienced an unprecedented credibility crisis (Ioannidis, 2005; John et al., 2012). The beginning of this crisis could be considered when Ioannidis (2005) published a paper in 2005 called "Why most published research findings are

false” in which he claimed that most research findings are fragile, unreliable and cannot be replicated. This increased the awareness of problematic issues affecting scientific literature and brought scientist’s attention to research findings. Consequently, many replication studies have been conducted in order to check whether or not the same outcomes could be found by repeating the same procedure. However, replication studies’ results have not found effects as strong as the original studies and in some cases, they have not found any effect.

According to Ioannidis (2005), the high rate of unreproducibility of research findings is the consequence “of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05” (Ioannidis, 2005, p. 696). In fact, p -value can’t be the only factor considered to interpret a research. Moreover, there are some practice that can inflate the probability of findings a p -value that is below the critical value, ultimately undermining the validity of a study’ conclusions, such as “Questionable Research Practices” (QRPs), “Questionable Measurement Practices” (QMPs) and low statistical power.

Even though this article was published in a medical journal, truth is that similar concerns have been found across the scientific landscape. Specifically, this crisis is most pronounced within psychology (Pashler & Wagenmakers, 2012). In fact, many articles have been published since 2011 that have shown how often psychologist are reluctant to share their published data for reanalysis (Wicherts et al., 2011) and how easily they engage in questionable research practices (QRPs) in order to obtain a statistically significance result such as performing multiple analyses on the dataset and reporting only those that yield a statistical significance (Simmons et al., 2011). Many researchers even admitted to engaging in QRPs (John et al., 2012).

1.1.1 Questionable Research Practices

The lack of replication studies increased the awareness of problems in the published literature, especially in psychology. Whereas some reasons are inherently related to psychology as a science (Chambers, 2019), there are other aspects of the research process that could undermine the conclusions of a study. One of these aspects is the so-called “Questionable Research Practices” (QRPs).

QRPs is a wide term that defined “practices that exploit ambiguities in what is acceptable for the purpose of obtaining a desired result” (Flake & Fried, 2020). It is used to describe both unethical and ambiguous practices. Therefore, it can range from lack of transparency to practices arising from ignorance and negligence to fraud (Flake & Fried, 2020). Given the hypothetical-deductive (H-D) model of the scientific methods (see Figure 1.1), to which psychology adheres, there is a range of Questionable Research Practices that can compromised it, all of which contribute to create findings which might not replicated.

P-hacking, also called *data dredging* or *data fishing*, is defined as “exploiting researcher degrees of freedom to generate statistical significance” (Chambers, 2019). Researcher degrees of freedom is part of the research process and it exists regardless of the authors’ intent to exploit it. It includes a number of decisions such as the exclusion of statistical outliers, which conditions should be combined and which ones compared, whether or not to gather other participants and which control variables should be considered (Chambers, 2019; Munafò et al., 2017). This flexibility has also been defined by Gelman & Loken (2013) as a “*garden of forking paths*,” because every decision may lead to a different pathways. Problems arises when authors, against the H-D methods, make analytic decision after inspecting their data, by trying out different analyses on a data-set and selecting the most desirable outcomes.

As a consequence, because of this exploratory behaviour, the likelihood of a findings

Figure 1.1: The Hypothetical-Deductive Model and its threat. Reprinted from *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*, by Chambers, 2019, Princeton University Press.

to be a false positive, usually set at 5% (i.e. $p < .05$), necessarily increase. To demonstrate how researcher degrees of freedom influence the false-positive rate, Simmons et al. (2011) simulated an experiment in which they tested the influence of four common degrees of freedom (i.e. choosing sample size and among dependent variables, using covariates and selecting subsets of experimental conditions), showing that if authors had used all of them, the false positive rate would have increased from 5% to almost 61%.

Furthermore, a key feature of *p*-hacking is that such decisions are hidden and never published. In fact, authors often portray all analysis as confirmatory (a priori) and hypothesis driven even though many were actually exploratory (post hoc). Whereas post hoc analyses could be useful tools for identifying potentially new research questions, they should be reported as such. Otherwise, readers will never be able to tell whether or not the results were selected among other non-significant outcomes (i.e. *cherry picking*),

and consequently they wouldn't be able to evaluate the validity of a study's conclusion (Chambers, 2019).

Moreover, following this questionable practice, researches not only will be able to find a statistically significant result in almost every set of data (Chambers, 2019) (increasing, at the same time, the false-positive rate), but they will also be able to generate a novel product that has higher chances to be published in a prestigious journal, due to publication bias.

Publication bias is defined by Chambers (2019) as form of *confirmation bias*, that occurs in published academic research, and that favors positive results (i.e. statistically significance differences or associations between conditions) over null results. A reason why this happens is the drive to publish positive results. Journals regards positive findings as more interesting for readers and as a reflection of greater degree of scientific advance. On the contrary, null outcomes are considered as scientific failure and something that researchers should avoid. In fact, in an environment that consider publishing in a prestigious journals a "form of "winning" in the academic game of life" (Chambers, 2019, p. 22) researchers are push to design experiments in which outcomes of interest would always be positive.

Another cause of publication bias is the quest of novelty (Chambers, 2019). Many journals judge articles according to novelty, some even explicitly use it as a policy for publication. This leads authors to either adopt a novel methodology or produce novel outcomes, yet the problems with novelty is that no single study can claim to be a discovery, even more in psychology where most research findings are probabilistic and not deterministic. A result can be classify as a discovery only if it can be reproduced and it's replicated.

Publication bias also lead researchers to alter the hypotheses after inspecting the data to predict unexpected outcomes. This procedure is another questionable research practice called *HARKing* (Hypothesizing After Results are Known). The term "HARK-

ing” was coined by Kerr (1998) and it refers to instances in which “researchers covertly alter their predictions after data have been analyzed in the interest of narratives flair” (Chambers, 2019, p. 31). The most common practice is to reverse the prediction after the analysis. In this way, authors are able to pretend that they predicted results that were actually unexpected and to report research findings that seem reliable with previous published researches, increasing the likelihood that the study will be published.

1.1.2 Questionable Measurement Practices

The research method requests authors to make decisions in all phases of the process and a fundamental part of it is considering how to define and measure what is being studied. In psychology, this step could be difficult. In fact, it’s rare that a construct such as depression or anxiety has an universally accepted definition due to the inability to direct observation. Moreover, it is hard to find a psychological construct that has one rigorously validated measure, with no degrees of freedom regarding how to score it. For instance, according to Santor et al. (2006) depression can be measured with at least 280 different scales.

Therefore, reporting transparently detail information about the measures is necessary to enable the scientific community to evaluate the validity of a study’s conclusion. However, despite its fundamental role, many published research do not provide information regarding measurement (Flake & Fried, 2020). As reported by Cybulski et al. (2016) about 1 in 3 trial registrations published in psychology journals did not provide such information.

When researchers “make decision that raise doubts about the validity of the measures, and ultimately the validity of study conclusions” (Flake & Fried, 2020, p. 458), it is referred to as “Questionable Measurement Practices” (QMPs). Neither rigorous research design, nor large sample, nor advanced statistic can be a solution to correct such false inferences. Like QRPs, QMPs can go from accidental omission of information

to practices meant to mislead and misrepresent, but, in this case, a lack of information introduces uncertainty in all aspects of the study's validity.

According to (Cook et al., 2002), there are four type of validity and they all contribute to the overall validity of a conclusion: internal validity, external validity, statistical-conclusion validity and construct validity. *Internal validity* includes all those aspects of a study's design that support causal claims about the relations between variables. Lack of information about the measures prevents researchers from properly evaluate if the measurement properties differ between two treatment conditions or across time. *External validity* allows authors to establish if the outcomes can be generalize across different populations and settings. It can be threaten by the lack of information regarding whether or not measures are sample or population specific. *Statistical-conclusion validity* concerns whether outcomes from a statistical analysis are correct. This kind of validity is more related to QRPs. In the end, *construct validity* refers to how the constructs in a study are operationalized.

Each type of validity can be evaluate only if precise information regarding measurement are provided. That is why transparency is fundamental. Moreover, measurement error "adds noise to predictions, increases uncertainty in parameter estimates, and makes it more difficult to discover new phenomena or to distinguish among competing theories" (Loken & Gelman, 2017, p. 584) and it can contribute to inflate the estimates of the effect. Thus, finding a statistically significance result under noisy conditions could lead to misleading conclusions.

Therefore, Flake & Fried (2020) proposed a series of questions (see Figure 1.2) that could help researchers to both avoid questionable measurement practices and facilitate the evaluation of researches' validity. Also, answering these questions provides enough information to enable a meaningful replication of the study.

Question	Information to report
1. What is your construct?	Define the construct Describe theories and research supporting the construct
2. Why and how did you select your measure?	Justify the measure selection Report existing validity evidence
3. What measure did you use to operationalize the construct?	Describe the measure and administration procedure Match the measure to the construct
4. How did you quantify your measure?	Describe response coding and transformation Report the items or stimuli included in each score Describe the calculation of scores Describe all conducted (e.g., psychometric) analyses
5. Did you modify the scale? And if so, how and why?	Describe any modifications Indicate if modifications occurred before or after data collection Provide justification for modifications
6. Did you create a measure on the fly?	Justify why you did not use an existing measure Report all measurement details for the new measure Describe all available validity evidence; if there is no evidence, report that

Figure 1.2: Questions to Promote Transparent Reporting of Measurement practices. Reprinted from measurement schmeasurement: questionable measurement practices and how to avoid them, by Flakes, J.K., Fried, E.I., 2020, *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465.

1.1.3 Low Statistical power

The low replication rate is not caused just by Questionable Research Practices and Questionable Measurement Practices. Another factor that should be taken into consideration and that could partially explain the replication crisis in psychology is the low statistical power.

Under the Neyman-Person approach, that requires to define both the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1), statistical power ($1 - \beta$) is defined as the probability of correctly rejecting the Null Hypothesis, when it's false. The inferential risks related to this approach are two: Type I error, defined as the probability of rejecting H_0 when it is true, also refers to as α ; Type II error, defined as the probability of not rejecting H_0 when it's false, also called β . For instance, if β is set at 20%, it means researchers accept the risk to incorrectly reject H_1 20% of the times assuming to repeat the experiment an infinite number of times.

It's now widely known that psychological studies are underpower (Bakker et al., 2020; Ioannidis, 2005). Researchers have been aware of this issue since 1962, when Cohen

surveyed all studies published in 1960 in the *Journal of Abnormal and Social Psychology* and showed that said studies had only 18% and 48% power to detect small and medium effect size, respectively; even large effect study had a 17% chance of being missed (Chambers, 2019; Sedlmeier & Gigerenzer, 1989). However, despite these findings, no changes occurred in the following years. In fact, when Sedlmeier & Gigerenzer (1989) surveyed the articles published in 1984 in the same journal, they found that small, medium and large effect studies had respectively a 16%, 46% and 89% chance of being detected. And more recently, Button et al. (2013) uncovered an even lower average power of 20% in a sample of findings in neuroscience.

Given that the power of a study tells the likelihood of a statistical test to detect an effect if it truly exists, a consequence of underpower studies is that if the effect of interest actually exists, the study will have a low probability to find it (Altoè et al., 2020; Bertoldo, 2019). This leads to a high rates of false negative that reduces the reliability of psychological research and, combined with the pressure for authors to produce positive results, can ultimately cause damage to theory generation (Chambers, 2019). In fact, Null results are more likely to be discarded and forgotten due to publication bias.

Moreover, due to this lack of sensitivity, any effects revealed by an underpower study must be large in order to obtain a statistically significant results (Chambers, 2019). However, the risk is that if a statistical significance is achieved in this condition, the effect size might be “too big to be true” (Button et al., 2013; Gelman & Carlin, 2014). Researchers may even assume that if they managed to achieve a statistically significance results in an underpower study, then the result is even more remarkable. Though, this is a mistake called by Loken & Gelman (2017) the “what does not kill statistical significance makes it stronger fallacy.” In fact, the statistical significance could be the results of other factors rather than the presence of the true effect. For instance, researcher degrees of freedom, sampling and measurement errors and extraneous variables all contribute to inflate the probability of findings a p -value that is below

the critical value (Altoè et al., 2020; Bertoldo, 2019).

Further, in these cases, what could be interpreted as a victory is in fact a loss (Altoè et al., 2020). Truth is that despite the statistical significance, what researchers actually found was not a discovery, but rather an inflated estimate of that effect (Button et al., 2013). This error is referred to as the “Winner’s curse.” In the simulation presented in Figure 1.3, in order to obtain a statistically significant result in a research design where the plausible true effect size is small, e.g. Cohen’s d of 0.20, and where researchers intent to perform a two-tailed test to compare two independent groups with 33 participants each, authors have to overestimate the true effect size. Otherwise, they won’t be able to reject the Null Hypothesis and find an effect close to what is considered “plausible” (i.e. 0.20), at the same time (Altoè et al., 2020). As it can be seen in Figure 1.3, in fact, due to the low statistical power (i.e. the actual power is 13%), an effect size of 0.2 can be found only outside the “rejection region” of H_0 ; while in the “rejection region” of H_0 (i.e. outside the vertical black lines) can be found effects that are either higher than 0.49 or smaller than -0.49. Clarifications regarding the concept of “plausible effect size” will be provided in the next chapter.

Figure 1.3: The Winner’s Curse. Reprinted from *Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis*, by Altoè, G. et al., 2020, *Frontiers in psychology*, 10, 2893.

Therefore, taking into account the statistical power of a study, as well as other important information, is fundamental to prevent completely misleading results (Altoè et al., 2020). Thus, Gelman & Carlin (2014) proposed two more type of errors in addition to Type I error and Type II error to help authors evaluate the results of a study more fully. These errors are called Type M and Type S errors and they will be illustrated in detail in Chapter 2.

Finally, another misconception is to consider the probability of those positive results to be true necessarily higher (Chambers, 2019). It's important to not confuse the probability of data under the null hypothesis (p) with the probability that a statistical significant result is a true positive. The latter must be estimated through the positive predictive value (PPV), which is calculated by the following formula, where R represents the prior probability that the effect being studied is truly positive:

$$\text{Positive predictive value PPV} = \frac{(1 - \beta) \times R}{(1 - \beta) \times R + \alpha \times (1 - R)}$$

As it showed, the PPV is directly related to statistical power: when power increases, so does the probability that an obtained positive result is a true positive (see Figure 1.4). Thus, high statistical power is necessary not only to reduce the rate of false negatives, but also to achieve a higher PPV.

1.2 Replicability and Reproducibility

To better understand what we mean when we say that a study was successfully replicated or did not replicated, it is important to have a definition of replicability and how it differs from reproducibility. In the literature, there isn't a standard definitions for these terms (National Academies of Sciences, Engineering, and Medicine, 2019): in many fields, reproducibility and replicability have distinct meanings, but different communities used opposing definitions (Peng et al., 2006); in other cases, qualifying

Figure 1.4: The relationship between Statistical Power and the Positive Predictive Value: when the Statistical Power increase, so does the probability that an obtained positive result is a true positive. Reprinted from *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*, by Chambers, C.D., 2019, Princeton University Press.

terms have been added to the definition, such as “method reproducibility,” “results reproducibility” and “inferential reproducibility” (Goodman et al., 2016). Consequently, the assessing of replicability and reproducibility is often difficult and researchers should always be prudent when interpreting the outcomes of a replication study.

The National Academies of Sciences, Engineering, and Medicine (2019) defined *reproducibility* as instances in which original study’s data and codes are used to recreate the results. Therefore, reproducibility is related strictly to transparency, because it “depends only on whether the methods of the computational analysis were transparently and accurately reported” (National Academies of Sciences, Engineering, and Medicine, 2019, p. 45). As a result, a reproducible study is a study that can be checked since the data, code, and methods of analysis are accessible to other researchers.

On the other hand, replicability is often defined by common understanding as a study in which a research’s procedure is repeated to observe whether the original findings recurs. However, this definition is incorrect (Nosek & Errington, 2020). In fact, if the

focus is on the technical method and replication requires the manipulated and measured events, then it's not possible to replicate an observational research or a research on past event: a questionnaire can't be administered in Italy and in the United States without taking into account the linguistic and cultural differences; at the same time, it's not possible to replicate exactly a phenomenon or an historical event, as an election. Thus, Nosek & Errington (2020) proposed a definition for replication in which the focus is on the role of replication and on the interpretation of possible outcomes.

Consequently, *replicability* is defined as a “study for which any outcome would be considered diagnostic evidence about a claim from prior research” (Nosek & Errington, 2020, pp.). Following this definition, a study can be labeled as a replication only if both this condition are met: results that are compatible with a prior claim improve trust in the claim, and outcomes that are inconsistent with a prior claim decrease confidence in the claim. If this criteria are fulfilled, replication becomes a mechanism for confronting existing understandings with new evidence. In this way, successful replications increase confidence in existing models while unsuccessful replications decrease confidence and encourage theoretical innovation in order to strengthen or replace a theoris.

Designing a replication with different methodology is not easy and requires deep knowledge of models and methods. However, define replication as a confrontation of existing evidences highlights its key role in scientific progress: a single study never definitively confirm or disconfirm a theories because only evidence of replicability using new data can establish the credibility of scientific claims (Schmidt, 2009).

Exact or direct replication is defined as a research in which the same results are achieved by following the same procedure measurement, and analyses of the original study (Goodman et al., 2016) and it is based on the assumption that a phenomenon occurs only in the same condition to the original. No such thing exists (Iso-Ahola, 2020). Moreover, exact replication neither can be the final arbiters of scientific truth nor can offer categorical answers to whether or not something exists (Iso-Ahola, 2020).

As explained above, in fact, it is not possible to recreate the exact condition to the original test and failures to replicate procedure, result in failure to replicate outcomes. This is especially true in psychology, where phenomena are not restricted to a particular condition or form, but can be seen in different situation. Furthermore, according to Iso-Ahola and Seppo (2020), problems of unreliability, invalidity and sampling errors, make methods psychometrically inadequate to provide yes-no answers. As a consequence, all attempts at direct replications will fail, because they try to replicate something that is tightly related to a specific time and specific context.

Conversely, *conceptual replication* is defined as a study that use different methods to test the same hypothesis as a prior study (Nosek & Errington, 2020). It is useful to highlight theories' strengths and weaknesses, as well as boundary conditions. Therefore, it can be an important tool to modify or expand theories and improve estimates of the effects (Iso-Ahola, 2020). However, many studies with this label are not replication by the definition given above. In fact, this studies are not designed to "decrease the confidence in prior claim if the outcome are inconsistent whit the claim." At most, failures are interpreted as identifying boundary conditions. Consequently, this studies can't be classified as replication, but rather as generalizability tests.

A *generalizability test* is a test in which many conditions might support the claim because of theoretical immaturity and unclear expectation, but failures do not decrease the confidence in the prior claim (Nosek & Errington, 2020). In this prospective, a replication is a subset of a generalizability tests. Thus, every replication study assesses generalizability, but not every generalizability test is a replication. Improvement of theoretical expectations help to clarify when replicability is expected, to the point where a complete model eliminates generalizability because theoretical expectations and future predictions are so precise that every study is a replication study. On the contrary, repeated failures reduce both generalizability and replicability, eventually to a model so weak that it makes no commitments to replicability.

1.3 Pre-registrations and Open Science Framework

The credibility crisis increased the awareness of problematic issues in published literature and promoted the creation of new practices which would guarantee rigour, reproducibility, and transparency. These aspects are considered a “set of gold standard” that has been proven to be fundamental for discovery (Chambers et al., 2014), and they should be used to judge the quality of a study, rather than novelty and positive results. In this way, researchers wouldn't be motivated to engage in questionable research practices with the aim of reporting a “publishable” study.

Different solutions have been proposed to solve this problem. One possible answer is the creation of online long-term data repositories, like the Open Science Framework (OSF). The OSF “is a tool that promotes open, centralized workflows by enabling capture of different aspects and products of the research lifecycle, including developing a research idea, designing a study, storing and analyzing collected data, and writing and publishing reports or papers” (Foster & Deardorff, 2017, p. 203). It was developed by the Center for Open Science (COS) in order to encourage openness, integrity, and reproducibility in the scientific researches.

One of the main features of the Open Science Framework is the possibility to register and pre-register a study. Specifically, a pre-registration is defined as a time-stamped, read-only version of a research plan created before collecting the data, that contains information regarding the hypotheses, the data collection procedures, the manipulated and measured variables, the statistical model and the inference criteria. Moreover, it helps to clearly separate confirmatory analyses from exploratory analyses, preventing both *p*-hacking and HARKing. Finally, a link to the pre-registration is requested to be included in the final manuscript.

A particular format of pre-registration is Registered Report (RR). In a RR, the manuscript undergoes through a two-stage peer review process (see Figure 1.5): prior

collecting the data (Stage 1) and after the results are known (stage 2). Research designs that pass Stage 1 peer review are offered ‘in principle acceptance’ (IPA) that guarantees publication of the manuscript regardless of the outcomes, as long as the methodology is carried out in accordance with the approved protocol. In fact, RR are based on the assumption that the quality of science should be evaluated depending on the importance of the research question and the rigor of the methodology, rather than whether or not positive outcomes are found (Chambers, 2019). After the data collection, a Stage 2 manuscript is resubmitted. When this process is complete, the final article is published.

Figure 1.5: Registered Report Protocol. Reprinted from ‘Center for Open Science’, 2019, <https://cos.io/rr/>

This format is designed to reward best practice and to prevent Questionable Research Practices. Moreover, it requires authors to adhere to the hypothetico-deductive model by think carefully about all phases of the research process and about some aspects that sometimes are ignored or forgotten, such as whether or not issue could arise and how to solve them and if the claims they want to make have the capability of being falsified (Chambers, 2019; Lakens, 2019). In fact, planning the research project carefully is crucial to be able to make accurate decisions regarding, for example, the sample size necessary to detect an effect size of interest (Bertoldo, 2019). The key rule of planning a research project will be examined further in the next chapter, when will be described the Design Analysis.

1.3.1 Aims

Psychology's credibility crisis had some beneficial side-effect. In fact, in order to find a solution to the low replication rate, scientist worked and are working to refine scientific standard and improve psychology as a science. Some contributions have already been illustrated, such as pre-registration and the Open Science Framework, while others will be presented in the following chapters (e.g., Design Analysis).

In the present work, the aim is to proposed a new statistical method that can help researchers to explore the inferential risks associated with each study in a meta-analysis. Specifically, this method will be based on Design Analysis, a power analysis approach developed by Gelman & Carlin (2014) to evaluate two more inferential risks in addition to Type I and Type II errors, namely Type M error, also called Exaggeration Ratio (i.e., the expected average overestimation of an effect that emerges as statistically significant) and Type S error (i.e., the probability of finding a statistically significant result in the opposite sign of the plausible true effect size).

The aim is mainly exploratory. This method, in fact, will allow researchers to evaluate the results of each study in the meta-analysis and have a clearer picture of the meta-analysis as a whole, at the same time. In this way, researchers will be able to make more conscious and statistically reasonable decisions during the research process.

Chapter 2

Design Analysis

In this chapter, we will analyze the inferential risks that researchers are exposed to when evaluating study results from underpowered studies using thresholds based on statistical significance. In these circumstances, researchers could either find an exaggerated estimate of the effect or a statistically significant result in the wrong direction or both. This, combined with the presence of publication bias, can explain why effects in the literature are biased upwards and replication studies mainly find smaller effect sizes. In order to estimate these risks, researchers can perform a *design analysis* which allows considering the probabilities of a Type M (magnitude) error and Type S (sign) error in a study. Both these errors will be define and their relationship with statistical power and sample size will be analyze. The chapter will conclude with a practical example of design analysis.

2.1 The design analysis

As discussed in the previous chapter, one of the possible explanations for psychology's credibility crisis is that many psychological studies are underpower. This, combined with the tendency of researchers to rely solely on statistical significance as a thresholds to make decision, could lead to misleading results. In fact, the risk of finding an

exaggerated estimates of the effect size increase in an underpower study, due to the need to pass such thresholds.

Therefore, Gelman & Carlin (2014) expanded the definition of power analysis, which in general refers to the “determination of an appropriate sample size, at pre-specified levels of *Type I* and *Type II* errors and a “plausible effects size” (Altoè et al., 2020, p. 2), with the addition of two more types of errors: *Type M* (Magnitude) error and *Type S* (Sign) error. *Type M* error, also known as *Exaggeration Ratio*, shows the predictable average overestimation of an effect that emerges as statistically significant. On the other hand, *Type S* error represents the probability that a statistically significant result has the opposite sign of the plausible true effect size. This analysis is called *Design Analysis*, precisely to emphasize the importance of both the study design (Altoè et al., 2020; Bertoldo et al., 2020; Gelman & Carlin, 2014) and the role of estimates and uncertainties in making more reasonable statistical claims (Altoè et al., 2020; Gelman & Carlin, 2014).

In fact, even though design analysis should be ideally performed when designing a study, as a simple size planning strategy (i.e., *prospective design analysis*), it can also be efficiently used to evaluate the already obtained study results and the associated inferential risk (i.e., *retrospective design analysis*) (Altoè et al., 2020; Bertoldo et al., 2020). Furthermore, design analysis reserves particular attention to the identification of the plausible magnitude and direction of the effect under study, refers to as *plausible effect size*. Both errors are indeed defined starting from its formalization (Bertoldo et al., 2020; Gelman & Carlin, 2014).

Even though the concept of plausible effect size will be examined further in this chapter, it is important to clarify that this work won’t adopt the effect size Gelman & Carlin (2014) used to develop the design analysis, which was an unstandardized measure (i.e., the difference between two means). In fact, in psychology literature, there are currently two adaption of Gelman & Carlin (2014)’s method: Altoè et al.

(2020) and Bertoldo et al. (2020). Both this contribution have the merit to emphasize the possible application of design analysis in psychology, considering standardized effect size measure (i.e. Cohen's d and *Pearson Correlation Coefficient*, respectively), which are more common in psychological studies. Therefore, the method we propose will not be based on an unstandardized measure of the effect size, but rather on correlation coefficient.

It should also be noted that the design analysis has been analytically formalized by Lu et al. (2019) and that it is currently available an R package for Prospective and Retrospective Design Analysis called *PRDA* developed by Callegher et al. (2021), that allows researchers to evaluate inferential risks (i.e., statistical power, Type M and Type S error) in a study considering a standardized measure of effect size. More information about the package are available at <https://claudiozandonella.github.io/PRDA/>, while the package itself is available both from GitHub (<https://github.com/ClaudioZandonella/PRDA>) and CRAN (<https://CRAN.R-project.org/package=PRDA>).

2.1.1 Type M error, Type S error and Statistical Power

Underpower studies are more likely to find an exaggerated estimates of the effect and to carry the wrong sign. As aforementioned, Gelman & Carlin (2014) formalized two indexes to quantify these risks: Type M (Magnitude) error and Type S (Sign)error.

Type M error is defined as the predictable average overestimation of an effect that emerges as statistically significant. It is important to note that this inferential risk is not a probability but rather a ratio between the absolute value of the mean of all the effect sizes that are statistically significant and the plausible true effect size. And it indicates the average percentage of inflation (Bertoldo, 2019). Consequently, if there is no exaggeration of the effect that emerges as statistically significant, the value that is obtained is one (i.e., minimum obtainable value). Ideally, Type M error should be as

close to one as possible to increase the probability of finding an accurate estimate of the effect.

Type S error is the probability of finding a statistically significant result in the opposite direction to the plausible one. It is calculated by making the ratio between the probability to find a negative statistically effect size and the power of a test, when the plausible effect size is positive. Otherwise, the ratio is between the power of a test and the probability to find a positive statistically significant effect (Bertoldo, 2019).

Furthermore, both this errors are strictly related to power. As it can be seen in Figure 2.1, the lowest the power, the higher the probability of committing a Type S error or a Type M error. However, Type M error increases exponentially as soon as the power goes below 50%, while it takes till the power is lower than 20% for Type S error to increase exponentially.

Figure 2.1: Relationship between power, Type S error and Type M error. Reprinted from *Beyond Power Calculation: Assessing Type S and Type M Errors*, by Gelman, A. et al, 2004, *Perspectives on Psychological Science*, 9, 6, 644.

Therefore, high level of power are required to minimize Type M and Type S errors. But, what does statistical power depend on? “In any experimental design, power depends on sample size, measurement variance, the number of comparisons being performed, and the size of the effects being studied” (Gelman & Carlin, 2014, p. 641). Though, in general, sample size is the factor that receive more attention. In Figure 2.2 it can be seen as power and sample size are directly related, meaning that when sample

size increases so does statistical power. In this simulation, with an independent sample t-test, a plausible effect size set to a Cohen's d of 0.35 and a Type I error established at alpha of 0.05 and, a power of 0.8 can be reached with about 130 subjects per group. Further, it also can be noticed that both Type M and Type S errors decrease when sample size increases. In fact, with 130 participants per group, Type M error is close to 1, whereas with less than 50 subjects not only the exaggeration ratio reach value higher than 1.5, but also the Type S error becomes worrisome.

Figure 2.2: Relationship between sample size and Power, Type M, and Type S. Reprinted from *Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis*, by Altoè, G. et al., 2020, *Frontiers in psychology*, 10, 2893.

2.1.2 Plausible effects size

Type M and Type S error are defined starting from the formalization of the “plausible effect size.” Providing an estimate of the effect size before running an experiment might seem paradoxical, especially given that the aim of the study is exactly to find what that estimate is. However, studies are rarely developed in a void. On the contrary, hypothesis are derived from theories that make predictions, with a different degree of precision.

In an epistemological perspective, strong theories are those that make detailed predictions and could be theoretically more easily disconfirmed, due to the presence of more

observations allowed to disconfirm the theory (Altoè et al., 2020; Kruschke, 2013). On the other hand, weak theories make larger predictions, therefore they are more easily confirmed because the only condition allows to disconfirm them it's the observation of no correlation. For instance, a strong theory could hypothesize a medium-sized positive correlation between two variables, whereas weak theories could hypothesize a correlation between two variables without specifying either the strength or the direction of the correlation (Altoè et al., 2020; Dienes, 2008). According to Karl Popper (1902–1994), theories that explain virtually everything and are hard to disconfirm risk to be out of the realm of science (Altoè et al., 2020; Dienes, 2008). Hence, scientific theories should provide at least a hint regarding the effect that is expected to be observed (Altoè et al., 2020)). But, how do you establish the dimension of this effect?

In general, authors take the estimate found in a pilot study or from a previous study published in the literature as a “plausible effect size.” However, this practice might be undesirable. In fact, the plausible effect size is defined as “an assumption the researchers make about which is the expected effect size in the population” (Bertoldo et al., 2020, p. 2). Therefore, its estimates should not be based on the noisy results of pilots or single studies but rather on the evaluation of the literature. For example, Gelman & Carlin (2014) suggest to use theoretical and literature reviews and meta-analysis. Though, when choosing this procedure, it's important to take into account the presence of publication bias (Franco 204, altoè 2020)(Altoè et al., 2020). Other possibilities could be to take into account multiple values to evaluate the results in different scenarios (Bertoldo et al., 2020), to formalize the expert knowledge (Altoè et al., 2020) or to define the smallest effect size of interest (SESOI) (Flake & Fried, 2020). Whatever procedure is chosen, all assumptions that led to the identification of a plausible effect size must be communicated in a transparent manner. This allows to increase the information provided by a study and ensured more informed and statistically reasonable claims about results, whether they are significant or not.

2.2 Prospective and retrospective design analysis

Design Analysis can be successfully conducted both prospectively and retrospectively (Gelman & Carlin, 2014). However, thanks to the growth in interest towards studies preregistration, prospective power analysis is more commonly performed. In fact, in the planning phase of a study, it's a useful tool to estimate the necessary sample size that allows to maximize power, to detect the effect size of interest and to minimize Type M and Type S errors (Bertoldo et al., 2020).

On the other hands, retrospective design analysis is performed after the results are known to analyze the actual power of the study, Type M and Type S error, helping the interpretation of statistically significant results (Bertoldo et al., 2020). In this case, it is important to note that the analysis is conducted using the effect size found in the study, and not the plausible effect size.

In the following paragraphs, it will be illustrated an example of prospective and retrospective design analysis performed by Altoè et al. (2020) in order to better understand the concept of Type M and Type S errors and the benefits of design analysis.

2.2.1 Prospective design analysis

In the planning phase of a study, prospective design analysis can be an useful tool considering researchers' need to take into account several factors while making a decision (e.g., the plausible effect size, the statistical power, the sample size and the Type I error). But, how does it help? And how do you perform it?

To demonstrate the benefits of prospective design analysis, Altoè et al. (2020) simulated an experiment in which the aim was to evaluate the differences between two treatments (i.e., G1- innovative treatment vs G2- traditional treatment) in improving a cognitive ability. Specifically, they simulated a case study where the researchers hypothesized a difference between the means of G1 and G2, and a plausible effect size

of Cohen's d of 0.25, considering the possible presence of publication bias. Then, in order to plan the sample size, they establishing an alpha of 0.05 and fixed the power at 0.80. Finally, they performed a design analysis to evaluate the Type M and Type S errors associated with their choices. The latter was calculated using the R function *design_analysis()*¹ that required two information: the hypothesized effect size and the statistical power. In the output, it can be found a summary information about the effect size and the statistical power as well as the required sample size to reach those values (i.e., $n = 252$ for each group), the probability of finding a statistically significant result in the wrong direction (i.e., Type S error = 0.00) and the Exaggeration ratio (i.e., Type M error = 1.13). A Type M error of 1.13 means that statistical significant results are on average overestimation of 13% of the plausible effect size.

Then, researchers decided to performed another design analysis, considering a lower level of power (i.e., power = 0.60), due to feasibility constraints in recruiting 504 participants. Consequently, they found that the required sample size was considerably smaller (i.e., $n = 158$ for each group), that the Type S error remained the same, and that the exaggeration ratio increased to 1.30.

At this point, researchers can evaluate which one is the best decision to make, considering the inferential risks associated to their choice. In fact, being aware of the implication of these risks allows researchers to communicate the results in a transparent and clear way, emphasizing the uncertainty related to their claims.

For instance, in this case, from a statistically point of view, the best choice would be the first scenario: the level of power (i.e., $d = 0.80$) is higher compare to the second scenario (i.e., $d = .60$) and Type M error is smaller (i.e., Type M = 1.13 compare to Type M = 1.30). However, in the decision making process, statistical factors should not be the only aspects considered, the phase of the study and feasibility constraints

¹*design_analysis()* is a R function created by @Altoè et al. (2020) which allows to perform both prospective and retrospective design analysis. In the first case, it requires to specify the plausible effect size (Cohen's d) and the power. On the other hand, retrospective analysis requires the sample size (n) and the effect size.

(as well as other factors) should also be taken into consideration. In fact, a sample size of 504 participants could be difficult to reach and could represent a constraint on feasibility.

2.2.2 Retrospective design analysis

In order to explain the benefits of retrospective design analysis, Altoè et al. (2020) adapted the case study presented beforehand and simulated three different, common scenarios that occur during the development of a study: (1) Relying on a single study published in the literature to estimate the sample size; (2) Feasibility constraints in recruiting the planned number of participants; (3) Prospective design analysis not performed due to the limited number of participants. Altoè et al. (2020) simulated the same results for the three scenarios, summarized in the following Figure 2.3.

Group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i> (<i>df</i>)	<i>p</i>	Cohen's <i>d</i> (95% CI)
Innovative treatment	31	114	16	3.496 (60)	0.001	0.90 (0.38–1.43)
Traditional treatment	31	100	15			

Figure 2.3: Comparison of the cognitive skill Y between the two groups. Reprinted from Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis, by Altoè, G. et al., 2020, Frontiers in psychology, 10, 2893.

Then, a retrospective design analysis was performed using the R function `design_analysis()`, with a Cohen's *d* of 0.25 (i.e., the plausible effect size) and a sample size of 31 subjects per group. The results showed a statistical power of 0.16, an average overestimation of 159% of the plausible effect size and a 1% risk of obtaining a significant result in the wrong direction.

In the first scenario,² where researchers relied on a single study that found a “big” effect size (e.g., 0.90) to estimate the sample size, retrospective design analysis shows

²This example was proposed to further emphasize the inferential risks related to the information provided by a single underpower study. However, as discussed in this chapter, it is strongly recommend to not plan the sample size based on a single study.

that if the plausible effect size were 0.25, the power of the study would be dangerously small (i.e., 16%) and the expected overestimation would be more than two times the plausible effect size (i.e., Type M = 2.59). Consequently, authors should not engage in the idea of estimate the sample size on a single published study. Indeed, they should examine more thoroughly theoretical and literature reviews to have a better knowledge of the phenomena and think about the opportunity to base the estimate of the effect size on the expert knowledge or meta-analysis (Altoè et al., 2020).

In the second scenario, after analyzing the output of the retrospective design analysis based on a plausible effect size of 0.25, researchers should be skeptical about their observed results. They should consider the opportunity to replicate the research on an independent sample, enlisting the assistance of additional experts in the field. Besides, they could also justify the need to recruit more participants thanks to the retrospective design analysis' results.

In the end, in the third scenario, researchers should present the results as descriptive, highlighting their uncertainty. Then, they could improve the study protocol, for example by increasing the reliability of the study variables, and if that is not possible, they should think about the possibility of abandon their research.

This examples, showed how performing a retrospective design analysis could help researchers in answering some important questions and issues that could arise after the research process has come to an end. For instance, a retrospective design analysis could help researchers in communicating the obtained result in a transparent way, in understanding the inferential risks associated with their decision and the consequences related to a low sample size or a small effect size of interest. However, it is also important to specified that design analysis is not a “automatic problem solver machine” (Altoè et al., 2020, p. 8). In fact, authors should always evaluate, case by case, what is a reasonable effect size and what are the optimal choices to make according to their research field.

Chapter 3

The Meta-Analysis

In the previous chapter, the context in which the statistic reform took place has been widely presented. In this chapter, another important method available to researchers will be presented: the Meta-Analysis. Initially a definition of Meta-Analysis will be provided, following the benefits associated with this techniques. Then the steps needed to conduct a Meta-Analysis will be presented, focusing mainly on the statistical analysis.

3.1 Meta-analysis

Psychology's credibility crisis and the resulting reform in the academic publishing, further highlighted the relevance of meta-analysis (Crocetti, 2016). In fact, in this context of statistic reform and given the shortcoming in relying solely on statistical significance as a thresholds to draw conclusions, effect sizes and confidence intervals have gained increasing importance in the decisions-making process about research findings. In fact, meta-analysis represents a robust method to gain more confidence in both the dimension of the effect being studied (i.e., effect sizes) and on the precision (i.e., confidence intervals) and additionally it provides an overall effect size and a confidence interval based on cumulative evidence from different studies (Cumming, 2013). However, before illustrating the benefits of this technique, it is important to clarify what is

a meta-analysis.

Meta-analysis has been defined by Glass (1976) as “the analysis of analyses”, in which the ultimate goal is to aggregate and contrast results from several related studies (Glass & Viechtbauer, 2010). Specifically, it refers to the use of statistical techniques to synthesize results across multiple primary studies (Hunt, 1997) or from studies selected with a systematic review. For clarity reasons, the latter refers to a qualitative research synthesis based on a clearly formulated question that uses systematic and explicit methods to identify, select, and evaluate a research in a specific area, and that provides an objective overview of all the evidence about the research question (Crocetti, 2016).

Conducting a systematic review with meta-analysis is considered to be the gold standard to reach a reliable synthesis of the results and the best way to get transparent and trustworthy conclusions in a field of interest (Crocetti, 2016). However, this technique can be conducted independently from each other, and in this paper, the focus will solely be on meta-analysis and how to perform it. More information regarding how to perform a systematic review with meta-analysis can be found in the article “Systematic Reviews With Meta-Analysis: Why, When, and How?” published by Crocetti (2016) in 2016.

3.2 The Benefits Of Conducting a Meta-Analysis

In the scientific process, reaching an overall understanding of a problem and identifying sources of variation in conclusions represents an important step which can be achieved through the synthesis of results across studies (Gurevitch et al., 2018). However, meta-analysis is an important tool with multiple applications whose purpose is not only to synthesize evidence on the effects of interventions or to inform policy or practice (Borenstein & Hedges, 2009), but it also plays a role in designing new studies, and has implications for what model should be used to analyze the data, what sensitivity analysis

should be undertaken and how to interpret research's findings. Moreover, conducting a meta-analysis has several benefits.

The first one, it is related to the strengths of the methodology (Crocetti, 2016). Meta-analysis in fact allows for greater transparency and replicability: on one hand, study selection and the weight assigned to each study need to be clearly reported in the final article, and on the other hand, the transparency of the procedure ensures that the results can be replicated. Additionally, it helps to handle the conclusions of a large number of studies with the same research questions (Bushman, 2001).

Secondly, meta-analysis allows researchers to analyze whether or not interventions or treatments are effective, as well as the factors that can enhance their effectiveness (Crocetti, 2016). Furthermore, it helps to individuate which factors are moderators and can explain differences in the magnitude of the effect being studied and which factors are accounted for inconsistent results in the published literature. At the same time, this technique provides information regarding those conditions that, despite a seemingly consistent literature, could explain the variance in the effect being observed.

Thirdly, the main strengths of meta-analysis is related to the credibility's crisis occurring in several statistic-dependent fields (Cumming, 2013). As previously discussed, in fact, meta-analysis provides a global effect size and a confidence interval based on findings of many study, which increases the confidence of the precision and the dimension of the effect being studied. Therefore, it's a robust method that promote an estimation thinking based on the effect sizes and confidence intervals, rather than a dichotomous thinking focused on significant versus non-significant results (Cumming, 2013). The latter, in fact, is related to a number of issues such as "what does not kill statistical significance makes it stronger fallacy" and the Winner's Curse (Loken & Gelman, 2017), which undermine any information regarding the practical relevance of the results and further highlight the gap between statistical significance and practical significance.

At the end, Borenstein & Hedges (2009) suggested others important benefits of conducting a meta-analysis compare to *narrative literature review articles*. Narrative reviews synthesize the state of the science of a specific topic in a theoretical and contextual point of view, without describing databases and methodological approaches used to conduct the review nor an evaluation of the published literature (Rother, 2007). The main merit of these reviews articles is to continuing education in a specific theme or field due to the up-to-date knowledge that they provides. However, narrative reviews do not offer an explanation of the results of any research in the context of all the other studies and do not provide a good mechanism to discuss the magnitude of the effect. Instead, in meta-analysis, if the effect size is consistent from study to study, the estimate of the effect is evaluated accurately and reported that it is robust in the studies included in the synthesis; on the other hand, if the effect size is inconsistent, the extent of the variance is quantify and the implications of such variance are taken into consideration. Moreover, the core of the analysis are the effect sizes, while in a narrative review the focus is on p-values and consequently, whether or not the effect is not zero. Meta-analysis also allows to evaluate the dispersion of effects, and distinguish between real dispersion and spurious dispersion.

3.3 How To Conduct a Meta-Analysis

In order to conduct a meta-analysis five fundamental step need to be followed (Borenstein & Hedges, 2009). (1) In the first step, all the pertinent studies on the topic of interest need to be found. (2) Then, the selection of the studies occurs. Generally, this step is based on pre-specified inclusion and exclusion criteria that determine the eligibility of studies to be included in the synthesis. (3) In the third step, researchers extract the relevant information from each study, such as sample size, effect size and relevant study-level variables. If quantitative data are not adequately provided, re-

searchers can either contact study author for obtaining the data that are missing or exclude that article from the meta-analysis. (4) Then, statistical analysis took place. Usually, this phase requires to build the dataset, summarize effect sizes, evaluate and explain heterogeneity of effect size across studies, assess the impact of publication bias and visualize and interpret results. (5) In the end, the results can be published.

All the phases presented in the fourth step will be now discussed and illustrated in detail. Additionally, a definition of effect size will be provided. In fact, as previously mentioned, the core of meta-analysis is indeed the effect sizes (not p-value) and the meta-analysis itself is formally described as a synthesis of K compatible effects, Y_i .¹

3.3.1 Calculating An Effect Size For Each Study

The effect size is a value which represents the strength of a relationship between two variables or, more precisely, the magnitude of the treatment effect. For example, effect size could reflect the impact of a treatment on risks of infection or the difference in test scores for males and females.

In a meta-analytic context, the effect size and the variance of each study need to be known or calculated. An “effect” could be almost any aggregate statistic of interest (e.g., mean, proportion). In practice, the so-called measures of effect-size are often based on standardized mean difference (e.g., Cohen’s d , Hedges’ g), binary data (e.g., *risk ratio*, *odds ratio*, and *risk difference*) and correlations (e.g., *Pearson’s correlations*, Fisher’s Z) and can be estimated through certain types of statistic (Borenstein & Hedges, 2009). For instance, Cohen’s d can be precisely calculated using means, standard deviations, and sample sizes. However, if those data are not available, researchers need to explore other options to compute the effect size that might be less accurate (e.g., using statistical significance obtained from the test conducted to compare the mean scores of the two groups of interest).

¹ Y_i represents the observed effect

When an effect size for each study is obtained, the effect size calculation is combined with its confidence interval (i.e., measure of precision), as well as its variance and standard error. All these information should be reported in a *forest plot* (see Figure 3.1) with the statistical significance of the effect size.

Figure 3.1: An example of forest plot: the effect size is represented by the position of the square and the confidence interval by the horizontal line. Reprinted from Systematic reviews with meta-analysis: Why, when, and how? by Crocetti, 2016, Emerging Adulthood, 4,1, 3–18.

3.3.2 Summarizing Effects: Fixed-Effect Versus Random-Effects Models

Once an effect size for each study is calculated, the effect sizes are used to compute a summary effect, which can be described as the weighted mean of the individual effects. Usually, the mechanism used to assign the weights is the inverse variance method, in which the weight of the study is calculated as the inverse of study variance (i.e., $W_i = \frac{1}{V_i}$).² However, this method and consequently the meaning of the summary effect, depends on the “assumptions about the distribution of effect sizes from which the studies were sampled” (Borenstein & Hedges, 2009, p. 6). Specifically, to perform this

² V is the variance of the study, which decreases as the variability of data decreases and sample size increases. Note that the weight of the study increases as the variance decreases

analysis, researchers can choose between two statistical model: the fixed-effect model or the random-effects model (Borenstein & Hedges, 2009).

In the fix-effect model, the assumption is that a true effect size exists and it is common to all the studies. This means that all factors that could influence the effect size are the same in all the studies, thus all the studies share the same true effect size. As consequence, all observed variability is due to random sampling error inherent in each study: assuming an infinite sample size, the sampling error would be zero and the observed effect for each study and the true effect would coincide. Moreover, under the fixed-effect model, while assigning a weight to each study (i.e., the inverse of that study's variance), the only source of variance considered is the within-study variance V_{Y_i} for study (i):

$$W_i = \frac{1}{V_{Y_i}}$$

The summary effect is consequently an estimate of the common effect size.

On the other hand, the assumption behind the random-effects model is that the true effects are normally distributed. Therefore, the aim of this approach is to estimate the mean of a distribution of effects, not one true effect assumed to be common to every study. Additionally, all these effect need to be represented in the summary effect because each of them provides information about a different effect size. Therefore, when assigning a weight to each study, it considers both the within-study variance V_{Y_i} and the between-studies variance T_2 :

$$W_i^* = \frac{1}{V_{Y_i}^*}$$

Consequently, as shown in Figure 3.2 and 3.3, study weight are more balanced in this approach than in the fixed-effect model: large studies are not given too much weight and small studies are not given very small weight. In the end, the summary effect is an

estimate of the mean of all relevant true effects.

Figure 3.2: A forest plot showing relative weight under the fixed-effect model Reprinted from Introduction to meta-analysis by M.Borenstein, L.V. Hedges, J.P. Higgins, and H.R. Rothstein, 2009, John Wiley and Sons.

Figure 3.3: A forest plot showing relative weights under the random-effect model Reprinted from Introduction to meta-analysis by M.Borenstein, L.V. Hedges, J.P. Higgins, and H.R. Rothstein, 2009, John Wiley and Sons

Generally, the random-effect model is more common in a meta-analyses context because it represents a conservative approach which allows to generalize the results of the meta-analysis beyond the studies included in the statistical synthesis. In fact, the assumption that the true effect size is the same in all the studies might be implausible in a meta-analysis, in which usually the assumption is that the studies share enough information that it makes sense to synthesize them. However, when the studies included

in the meta-analysis are five or less, both analyses should be conducted because in these case the between studies variance T_2 is less accurate (Borenstein & Hedges, 2009).

3.3.3 Transformation of effect-sizes: Fisher's z transformation

Since meta-analysis is based on normality, when the effect size is based on correlations, its distribution is not normally distributed, so it is preferable to normalize the effect sizes via appropriate transformations (Borenstein & Hedges, 2009). Moreover, it's better to not run meta-analyses on the correlation coefficient itself because the variance is strongly dependent on the correlation. Therefore, the correlation coefficients (r) are usually transformed using the Fisher's z transformation and all analyses are performed using the transformed values. Once the results are obtained, they are converted back to a correlation (r). More precisely, Fisher's z transformation, the variance and the standard error of z are given by the following formula:

$$z = 0.5 * \ln\left(\frac{1+r}{1-r}\right)$$

$$V_z = \frac{1}{n-3}$$

$$SE_z = \sqrt{V_z}$$

3.3.4 Heterogeneity Assessment

Another important aspect to take into consideration after an effect size for each study is calculated is how to identify and quantify the dispersion of effect sizes across study (the assumption under the random-effect model in fact is that the true effect size may vary from study to study) (Borenstein & Hedges, 2009). This is important because if the effect size is consistent, the focus will be on summary effect and the effect will be considered as robust from study to study. On the contrary, if the effect size is not consistent, the focus will be on the dispersion itself, not on the summary effect, even

though it might be still reported.

In order to identify and quantify heterogeneity, however, researchers need to take into consideration that the observed variation is actually spurious, which means that it includes both true variation in effect sizes and random error, and that heterogeneity regards only the variation in the true effect sizes (Borenstein & Hedges, 2009). Consequently, the observed variation need to be partitioned into real heterogeneity in effect size and the within-study error. The mechanism to do so requires at first to compute the total amount of study-to-study variation observed, then evaluate how much the observed effects would vary from each other if the true effect would be the same in all studies and finally if there is any excess variation, it is assumed to reflect real differences in effect size (heterogeneity).

In order to assess if the heterogeneity across studies is significant, Q statistic is computed (Borenstein & Hedges, 2009). Q is the weighted sum of squared differences between individual study effects and the summary effect.³ If the p -value associated with Q is significant, the null hypothesis of homogeneity can be rejected, meaning the true effects vary across study. However, a non-significant p -value should be not considered as an evidence of consistency in the effect sizes. In fact, low power, small number of studies and/ or large within-study variance could be responsible for the non-significant p -value.

Moreover, Q statistic assesses only the viability of the null hypothesis, not how large is the true dispersion. The latter can be estimate and described by computing different indices, such as the between-study variance T^2 , the between-study standard deviation T , the Higgins' I^2 (estimate, in percentage, of how much of the total variability in the effect size estimates can be attributed to heterogeneity among the true effects) and the estimate of the ratio of the total amount of variability in the observed outcomes to the amount of sampling variability H^2 . Large value of indices indicate a large unexplained

³Weights are those used in the meta-analysis.

heterogeneity (Borenstein & Hedges, 2009).

Generally, a small heterogeneity (i.e., I of value of 25%) means that results of primary studies are rather similar and consistent. However, it is important to note that it is common to find significant and large heterogeneity (i.e., I value of 75%) across study findings, especially when more than 10 studies are included in the meta-analysis and the main research question is addressed differently in each study. This usually happens due to variations in the effect size under study, rather than the statistical significance of the results (Crocetti, 2016).

3.3.5 Publication Bias Evaluation

Another important step while running a meta-analysis is to assess the presence of publication bias. As discussed in the first chapter, publication bias is a form of confirmation bias that favors positive results over null or negative results [Chambers (2019), and that consequently creates a systematic difference between published studies and unpublished studies (i.e., grey literature) (Rothstein & Hopewell, 2009). In the context of meta-analysis, publication bias means that the studies included in the synthesis differ systematically from all studies that should have been include. Consequently, it can make conclusions less trustworthy. Therefore, it's necessary to evaluate whether the studies included would change the results in a non-substantial way (i.e., moderate impact), would change the results in a substantial way (i.e., large impact) or would not change the results of meta-analysis (i.e., minimal impact).

The impact of publication bias can be assess with several methods, such as the Fail-safe N test (Rosenthal, 1979), the Egger's linear regression method (1997), Begg & Mazumdar (1994) rank correlation method, the p -Curve (Simonsohn et al., 2014) the trim and fill procedure (Duval, 2005).

Fail-safe N analysis is a common method used in the social sciences and it's computed when "the overall effect size is significant to know how many studies with a

non-significant result would be required to bring the combined effect size to be non-significant” (Crocetti, 2016, p. 12). Specifically, the assumption is that one can compute an effect size measure from the studies published and assess how many studies must be hidden in file drawers to outweigh the published results (Haaf, 2020). The main shortcomings of this method are that it does not consider negative results, assuming that in the grey literature results are on average null, nor it consider important information such as sample sizes and heterogeneity (Becker, 2005).

The *Egger’s linear regression method* (Egger et al., 1997) and *Begg & Mazumdar (1994)* both assess the asymmetry of a *funnel plot* (Light & Pillemer, 1984). The funnel plot represents “the effect size estimates from individual studies against some measure of each study’s size or precision” (Crocetti, 2016, pp. 13–14) (see Figure 3.4). Precisely, this graphs is based on the idea that the estimated effect size is a function of the sample size: with small samples is expected more variability in the effect sizes, while with larger samples is expected that the observed effect sizes would converge to the true effect size of the meta-analysis (Haaf, 2020).

Figure 3.4: Example of funnel plots: the circle represents the study, the observed effect size is reported in the x-axis and the standard error in the y-axis. In panel A publication bias is absent. Panel B shows an extreme publication bias. Reprinted from *Conventional Publication Bias Correction Methods*, by J.M. Haaf, 2020.

It appears symmetrical when publication bias is absent, while it resemble an asymmetrical (inverted) funnel, with a gap in a bottom corner of the graph, in the presence of publication bias.⁴ The latter condition, however, may also occur due to other reason, such as small study effects (Sterne et al., 2005). Therefore, to compensate for the interpretative difficulties of a funnel plot, the Egger's linear regression method and the Begg & Mazumdar (1994) rank correlation method have been developed. In both cases, when the results is statistically significant, it is indicative of a potential publication bias.

Another method to assess the publication bias is the p -curve (Simonsohn et al., 2014), in which the p -curve represent the distribution of the significant p -value in a meta-analysis. It allows to test whether or not a true underlying effect is present and to evaluate the average effect size. If the true underlying effect is larger than zero, the significant p -value should be distributed as a right-skewed (see Figure 3.5), otherwise if the true underlying effect is zero, the distribution should be uniform (Haaf, 2020).

Figure 3.5: P-curve graphs. The right-skewed shows that the effect size actually exists. Reprinted from Conventional Publication Bias Correction Methods, by J.M. Haaf, 2020.

A more recent method is the *trim and fill procedure*, defined as a non-parametric

⁴In the funnel plot, it assumes that the largest studies are plotted near the average, and smaller studies are spread evenly on both sides of the average, creating, in fact, a funnel-shaped distribution

data augmentation statistical technique that can be used to estimate the number of studies missing from a meta-analysis due to the suppression of the most extreme results on one side of the funnel [duvalsueandtwedierichardNonparametricTrimFill2000; Haaf (2020)]. In three iterative phases, this method allow to compute an adjusted effect size and its 95% confidence interval: at first, the asymmetrical part of the funnel plot is removed as affected by publication bias; then, the average effect size is estimated out of the remaining (symmetrical) studies; finally, the trimmed studies are re-included in the analysis and their symmetric counterparts are filled (see Figure 3.6. If the number of trimmed studies is zero or the difference between the observed and the adjusted effect size is trivial, then it means that there is no publication bias (Duval, 2005).

Figure 3.6: Trim and Fill procedure illustrated in a funnel plot. Black dots represents the original data set. Unfilled points show added data from the trim and fill procedure. The red line represents the resulting effect size estimate, closer to the real effect size (i.e., 0.5) then the estimate of the untreated dataset. Reprinted from Conventional Publication Bias Correction Methods, by J.M. Haaf, 2020.

Chapter 4

drmeta: A Function For Exploring Inferential Risks Associated With Each Study In a Meta-Analysis

In this chapter, our method to explore the inferential risks associated with each study in a meta-analysis will be presented. Specifically, this method allows researchers to conduct a retrospective design analysis on each study in a meta-analysis, in order to explore the inferential risks associated with each study (i.e., power, Type M and Type s errors). Initially, the new method will be introduced. Then, instruction and information on how to perform the analysis in R will be provided. Precisely, our function *drmeta* will be introduced. Firstly, an explanation of how the *drmeta* works will be provided, as well as the input required by the function and the output. In the end, an example of how to interpret the output will be also shown.

4.1 Introduction

During the research process, a difficult but crucial part is to establish what could be consider a plausible effect size, taking into consideration the research question and the field of interest (Altoè et al., 2020). Good practice should be to not rely on data at hand or on noisy results of a single study, instead the formalization of a plausible effect size should be based on other tools, such as extensive theoretical literature review, expert knowledge and/or on meta-analyses (Altoè et al., 2020).

In this paper, we will consider the latter case, in which a meta-analysis is used during the research process to estimate what could be consider a plausible effect size. Precisely, a new method is presented for exploring the inferential risks (i.e., power, Type M and Type S errors) associated whit each study in a meta-analysis, trough a retrospective design analysis.

The aim of this method is mainly exploratory. Firstly, results can be relevant both on an study level (to evaluate power, Type M and Type S errors for each study in the meta-analysis), and to get a clearer picture of the meta-analysis as a whole. Secondly, it allows to explore the inferential risks of each study whether the study is statistically significant or not, consequently going beyond the significant filtering. In the end, this method will allow researchers to explore the inferential risks in five different scenarios (“standard,” “Standard corrected,” “All studies,” “All studies corrected and”Personalized”).

Following what has been defined as the *standard method*, researchers are able to evaluate the inferential risks of each study in a meta-analysis considering all studies expect the one at hand. Oppositely, the *all studies* method considers all studies included in the meta-analysis, which may produces spurious results (because each study is included in the calculation of the effect-size and consequently in the computation of power, Type M and Type S errors), but at the same time could be helpful to get an overall estimate.

However, while conducting or considering a meta-analysis is always important to

remember to assess the impact of publication bias (see Chapter 1, paragraph 1.1.1), especially because meta-analysis is not a multi-laboratory study. In fact, even though on certain sites there are fixed and shares protocol to follow in order to conduct a research (e.g., Open Science Framework), not all studies published in the literature share the same protocol. Therefore, the variability can be high and in the context of a meta-analytic study, publication bias can be a risks that must be assessed. In our case, among the available methods, we have chosen to follow the trim and fill procedure (see Chapter 3, paragraphs 3.3.5), purely for illustrative purposes.

This is the reasons why both *standard* and *all studies* methods can be performed based on an adjusted effect size. Precisely, they are called *standard corrected method* and *all studies corrected method*.

In the end, the *personalized method* is available. This is a method that is not based on the meta-analysis itself. In fact, researcher is able to specify the effect size of interest and then explore the retrospective power of the study, based not on the effect size of the study, but indeed on the one chosen by the researcher. In this way, this method is similar to the researches which surveyed the power of already published reports to explore what is the actual power of the study (e.g., Cohen's 1962).

These analyses can be conduct in R trough our function *drmeta*. It is important to note that *drmeta* only works for correlation coefficient as a measure of the effect size. At fist, the effect sizes of each study are transformed into Fisher's z (see Chapater 3, paragraph 3.3.3) and then the overall effect size is back-transformed into correlation coefficient. However, the function can easily be adapted to other effect sizes measure.

4.2 *drmeta* Function: How It Works

drmeta is our new R function which allows researchers to explore the inferential risks associated with each study in a meta-analysis. Precisely, a retrospective design analysis

is conducted on a meta-analysis estimating the statistical power, Type M and Type S errors for each study included in the synthesis.

In order to reach this goal `drmeta` requires two packages to be installed: *metafor* and *PRDA*. The first one is a packages created by Viechtbauer (2010) to conduct a meta-analysis in R, while *PRDA* is an R package create by Callegher et al. (2021) for performing both prospective or retrospective design analysis (`drmeta` used only retrospective design analysis).

When both these packages are installed, the first step is to remove all objects¹ and set the work directory. Then, the function need to be loaded; generally, the command `source()` is used to achieve this goal. Note that the function `drmeta` must be present in the work directory, otherwise this command won't work. Now the function is ready to be used (the script of the function is available in the Appendix).

As previously discussed, `drmeta` allows five different methods:

1. *Standard*: for each study, the study is removed, redo the meta-analysis, and then on the synthetic effect are calculated power, type M and type S error (select option “*stand*”).
2. *Standard corrected*: similarly to the standard method, for each study, the study is removed, redo the meta-analysis, the synthetic effect is calculated and adjusted with the trim and fill procedure, then on the adjusted effect, power, type M and type S error are calculated (select option “*stand_c*”).
3. *All studies*: for each study, power, type M and type S errors are calculated based on the effect size of the meta-analysis considering all the studies included in the synthesis (select option “*all*”).
4. *All studies corrected*: similar to all studies, but in this case the effect size is adjusted with the trim and fill procedure (select option “*all_c*”).
5. *Personalized*: user must specify an effect size (select option “*pers*”).

¹usually the function `remove()` is used; `remove(list=ls())`

Researchers can choose to perform one or all of these method. For example, one could choose to use the *standard* method and then the *personalized* one to explore what changes. In fact, we advise to use this function in an exploratory way, in order make an informed decision based on the risks that might occur. Methods and their associated effect sizes are formalized in Table 4.1:

Table 4.1: The effect size considered in each method.

<i>Method</i>	<i>Effect Size</i>
Standard	Effect Size of the meta-analysis without the study considered
Standard corrected	Adjusted effect size of the meta-analysis without the study considered
All studies	Effect size of the meta-analysis considering all studies
All studies corrected	Adjusted effect size of the meta-analysis considering all studies
Personalized	Effect size defined by the researcher

4.2.1 drmeta Application: Input and Output

drmeta requires a dataset with three variables:

1. the name of the study
2. coefficient correlation (r)
3. sample size (n)

To better understand, supposed we have a dataset that for the sake of simplicity has six studies:

```
##      study    r    n
## 1 Study1 0.80  10
```

```
## 2 Study2 0.30 45
## 3 Study3 0.30 100
## 4 Study4 0.65 60
## 5 Study5 0.70 30
## 6 Study6 0.10 20
```

The name and the order of the variables must be as shown (i.e., *study*, *r* and *n*). Then, *drmeta* requires as input the name of the dataframe (*d*) and the method (*method*) which can be set as: “stand,” “stand_c,” “all,” “all_c,” and “pers.” Moreover, it requires to specify an effect size (*r*) in cases the chosen method is “pers.” Type I error rate (sig.level) is set as default at 5% and can be changed by the user.

Finally, the function *drmeta* returns a list of the main arguments and the results. More precisely, it shows the id (i.e., a progressive numerical code for each study), the name of the study, the coefficient correlations (*r*), the sample size (*n*) and then the power of the study, Type M and Type S error, “*r_meta*” and “*r_meta_cor*.” An example can be seen below. In this case, the dataset is the one created earlier (*dati*), the *standard* method is applied and the significant level is set 5%.

```
##   id  study    r    n power type_M type_S    r_meta r_meta_cor
## 1  1 Study1 0.80  10  0.79   1.22   0.03 0.4478691         NA
## 2  2 Study2 0.30  45  1.00   0.99   0.00 0.5307045         NA
## 3  3 Study3 0.30 100  1.00   1.00   0.00 0.5365004         NA
## 4  4 Study4 0.65  60  1.00   0.99   0.00 0.4428927         NA
## 5  5 Study5 0.70  30  0.97   1.01   0.00 0.4353529         NA
## 6  6 Study6 0.10  20  0.97   1.01   0.00 0.5392838         NA
```

Specifically, “*r_meta*” represents the synthetic effect study after the study has been removed, while “*r_meta_cor*” shows the effect size adjusted with the trim and fill procedure. For method 1, 3 and 5 (“standard,” “all studies” and “personalized”) the

column “*r_meta_cor*” will return *NA* due to the absence of the effect size correction. On the other hand, for method 2 and 4 (“*standard corrected*” and “*all studies corrected*”) will return the adjusted effect study.

In the example shown above, therefore, the “*r-meta*” shows the coefficient correlation of the meta-analysis excluding the study in question. For instance, 0.45 is the effect size of the meta-analysis without considering the study1. We can also observed how study1 has a risk of overestimating the plausible effect size of 22% and a Type S error of 3%.

Applying at the same dataset the “*standard corrected*” method, this are the results:

```
##   id  study    r    n power type_M type_S    r_meta r_meta_cor
## 1  1 Study1 0.80  10  0.26   1.65      0 0.4478691  0.4478691
## 2  2 Study2 0.30  45  0.97   1.01      0 0.5307045  0.5307045
## 3  3 Study3 0.30 100  1.00   1.00      0 0.5365004  0.5365004
## 4  4 Study4 0.65  60  0.86   1.07      0 0.4428927  0.3790303
## 5  5 Study5 0.70  30  0.70   1.17      0 0.4353529  0.4353529
## 6  6 Study6 0.10  20  0.72   1.13      0 0.5392838  0.5392838
```

Again, study1 has an high expected overestimation of the plausible effect size (i.e., 65%), with a small level of power, i.e., 27%.

Following the results of the “*all studies*” and “*all studies corrected*” method.

Applying the *All studies* method:

```
##   id  study    r    n power type_M type_S    r_meta r_meta_cor
## 1  1 Study1 0.80  10  0.32   1.52      0 0.4892741      NA
## 2  2 Study2 0.30  45  0.94   1.02      0 0.4892741      NA
## 3  3 Study3 0.30 100  1.00   1.00      0 0.4892741      NA
## 4  4 Study4 0.65  60  0.98   1.00      0 0.4892741      NA
## 5  5 Study5 0.70  30  0.81   1.09      0 0.4892741      NA
```

```
## 6 6 Study6 0.10 20 0.62 1.21 0 0.4892741 NA
```

Applying the *All studies corrected* method:

```
## id study r n power type_M type_S r_meta r_meta_cor
## 1 1 Study1 0.80 10 0.26 1.66 0 0.4892741 0.4473219
## 2 2 Study2 0.30 45 0.89 1.06 0 0.4892741 0.4473219
## 3 3 Study3 0.30 100 1.00 1.00 0 0.4892741 0.4473219
## 4 4 Study4 0.65 60 0.96 1.02 0 0.4892741 0.4473219
## 5 5 Study5 0.70 30 0.72 1.15 0 0.4892741 0.4473219
## 6 6 Study6 0.10 20 0.54 1.30 0 0.4892741 0.4473219
```

When both methods are performed, we can observed that the coefficient correlation of the meta-analysis (r_{meta}) remains the same. This is due to the fact that, as previously suggested, all studies are considered to run the analyses. In these cases, the results show that study1 has the highest Type M error (i.e., 52% and 65%, respectively) and the lowest level of power, i.e., 31% and 26%, respectively.

In the end, applying the “*personalized*” method a the same that set, selecting an effect size of 0.25, these are the results:

```
## id study r n power type_M type_S r_meta r_meta_cor
## 1 1 Study1 0.80 10 0.10 2.88 0.05 0.25 NA
## 2 2 Study2 0.30 45 0.39 1.54 0.00 0.25 NA
## 3 3 Study3 0.30 100 0.71 1.18 0.00 0.25 NA
## 4 4 Study4 0.65 60 0.49 1.38 0.00 0.25 NA
## 5 5 Study5 0.70 30 0.26 1.83 0.00 0.25 NA
## 6 6 Study6 0.10 20 0.19 2.16 0.01 0.25 NA
```

What we can observed is that study1 and study6 are remarkably underpower (i.e., 11% and 19%, respectively). Moreover, the expected overestimation (Type M error)

would be almost three times the plausible effect size (i.e., 190%) for study1 and two times the plausible effect size for study6 (i.e., 116%).

The potential of this method will be further explored in the case study in the next chapter.

Chapter 5

A Case Study: Exploring the potentials of *drmeta*

During the research process, hypothesizing a plausible effect size is an important step to take. Different tools are available to researchers to formalize it, one of them being the use meta-analysis. However, to make an informed and conscious decision, researchers should be aware of the inferential risks associated with the data they are considering and consulting. For this reason, we proposed a new method, illustrated in the previous chapter, to explore such risks: *drmeta*. In the following chapter, finally, we will illustrate the potentials of this function with a case study analyzing a meta-analysis published in the literature by Talò (2018).

5.1 The Original Study

In 2018 Talò (2018) published a meta-analysis called “Community-based determinants of community engagement: A meta-analysis research,” in which the social determinants associated with living in a community were analyzed. Specifically, the main community-based variables were identified and their impact on community engagement were calculated.

The dependent variable in this study is therefore the community engagement, defined “a process in which individuals take part in decision making in the institutions, programs, and environments that affect them” (Heller et al., 1989, p. 339; @ Talò, 2018) and “the active, voluntary involvement of individuals and groups in changing problematic conditions in communities and influencing the policies and programs that affect the quality of their lives and the lives of other residents” (Ohmer, 2007; Talò, 2018).

Then, from the analysis of the literature, the most frequent community-based variables associated with community engagement were identified. The variables considered were “variables that measure a certain relationship with one’s community” (Talò, 2018, p. 575), such as satisfaction, feeling of belonging and participation in the community. Therefor, since community engagement is distinct from conventional political participation, variables such a stand for election, collaborate in an election campaign, be a member of a political party, were not taken into consideration. A summary of the variables considered in this study is available in the table 5.1.

Table 5.1: Variables considered in the original study

Variables	Description
Date of publication	from 2003 to 2013
Country	USA, Spain, Italy, Iran, Japan, Peru, Asia, UK Cze Republic
Sample size	from 51 to 10.566
Percentage of female	from 21.1 to 68 per cent
Type of sample	Students, young, adult, elders
Mean age	from 19.6 to 74.41 years
Measuring instrument	ad hoc, one item, SCAP, Gracia et al (1996), CPS

Variables	Description
Community based predictors	Sense of community, social Life feeling, sociopolitical control, social support, community identity, social well-being, place Identity, trust in the community place dependence, ecological integrity economic dependency, intrapersonal empowerment trust in institutions,family identity, satisfaction, interactional empowerment

In the end, the meta-analysis were conducted only for predictors with at least two studies (see Table 5.2).

Table 5.2: Predictors with at least two studies

Predictors	<ul style="list-style-type: none">• sense of community• community identity• social well-being• place identity• trust in the community• trust in institutions
------------	---

5.2 Exploring The Inferential Risks Associated With Each Study In The Meta-Analysis: Application On A Published Study

In this section, we will explore the inferential risks associated with each studies in the meta-analysis presented previously. However, before conducting the analysis, we want to clarify that the following case study was chosen for illustrative purposes only and,

by no means our objective is to judge the study beyond illustrating an application of how *drmeta* works and how to use it on a published study.

To achieve these objectives, we selected one of the meta-analyses conducted in the article and based on that we will explore the inferential risks (i.e., power, Type M and Type S errors) associated with each included study. Specifically, we chose to investigate the association between the variables ‘sense of community’ and ‘community engagement.’ The reason behind this choice is that the ‘sense of community’ variable was included in the largest number of studies (i.e., 13 studies) compare to the other variables which were included only in 4, 3, or 2 studies. Specifically, in the latter cases, the application of the random-effect meta-analysis doesn’t often offer plausible estimate.

At this point, our dataframe contained 13 studies and 14 variables on the study level. However, for our purposes, it was not necessary to consider all variables included in the original study. Consequently, we chose to consider only those variables that are useful for showing how our function works (i.e., a progressive numerical code for each study (id), authors, year, correlation coefficient and sample size). Furthermore, we decided to eliminate the category “Elders” in order to have a more homogeneous sample. In the end, therefore, the dataframe on which we will conduct the analysis contains 12 studies and 5 variables (as shown in the table below).

##	authors	year	cor	n
## 1	Chang	2010	0.500	96
## 2	Chioneso and Brookins	2015	0.290	235
## 3	Christens and Lin	2014	0.090	1322
## 4	Cicognani et al. (IR)	2008	0.202	214
## 5	Cicognani et al. (IT)	2008	0.261	200
## 6	Cicognani et al. (USA)	2008	0.231	125
## 7	Mannarini et al.	2010	0.190	194
## 8	Mazzoni et al.	2014	0.245	835

```
## 9      Moreno-Jiménez et al. 2013 0.326 756
## 10 Ramos-Vidal and Maya Jariego 2014 0.010 120
## 11      Serek and Machackova 2015 0.190 573
## 12      Tsai et al 2008 0.516 51
```

Note that Talò (2018) included three studies by the same author (i.e., Cicognani) published in the same year (i.e., 2008). Generally, in this case, the meta-analysis should be conducted following a specific procedure in order to take it into account. However, with regard to our study, since the purpose is purely illustrative and the three studies are independent and conducted in three different states (i.e., Iran, Italy and USA), we considered them as three unique and distinct studies and we renamed them as “Cicognani et al. (IR),” “Cicognani et al. (IT)” and “Cicognani et al. (USA).”

In the next step we adjusted the dataframe so that *drmeta* could run (i.e., study, r , n), and we performed a meta-analysis with random effects in R on our dataframe. The results showed a significant effect size of small-medium dimension (i.e., $d = .243$, according to Cohen’s d definitions), with a confidence interval (IC) of 0.173-0.311. Then, applying the trim and fill procedure, we assessed the impact of publication bias. However, the results do not support the evidence of publication bias.

Finally, we explored the inferential risks associated with each study in this meta-analysis. Firstly, we decided to use the “*standard corrected*” method. In fact, even if the trim and fill procedure did not show the presence of publication bias, after removing a study, publication bias could still be detected. Therefore, we used the method with the correction.

```
##      id      study      r      n power type_M type_S      r_meta
## 1      1      Chang 0.500    96  0.48   1.43        0 0.2219230
## 2      2  Chioneso and Brookins 0.290  235  0.96   1.02        0 0.2395236
## 3      3  Christens and Lin 0.090 1322  1.00   1.00        0 0.2590618
```

## 4	4	Cicognani et al. (IR)	0.202	214	0.96	1.03	0 0.2479902
## 5	5	Cicognani et al. (IT)	0.261	200	0.94	1.04	0 0.2426163
## 6	6	Cicognani et al. (USA)	0.231	125	0.80	1.13	0 0.2450239
## 7	7	Mannarini et al.	0.190	194	0.94	1.03	0 0.2488134
## 8	8	Mazzoni et al.	0.245	835	1.00	1.00	0 0.2444694
## 9	9	Moreno-Jiménez et al.	0.326	756	1.00	1.00	0 0.2330830
## 10	10	Ramos-Vidal and Maya Jariego	0.010	120	0.82	1.11	0 0.2565792
## 11	11	Serek and Machackova	0.190	573	1.00	1.00	0 0.2501538
## 12	12	Tsai et al	0.516	51	0.37	1.59	0 0.2283708
##		<i>r_meta_cor</i>					
## 1			0.1918136				
## 2			0.2395236				
## 3			0.2590618				
## 4			0.2479902				
## 5			0.2426163				
## 6			0.2450239				
## 7			0.2488134				
## 8			0.2444694				
## 9			0.2330830				
## 10			0.2565792				
## 11			0.2501538				
## 12			0.2283708				

Taking a first look at the results, we can see that even when eliminating one study at a time, the effect size estimates remain homogeneous (see *r_meta_corr*, where the effect size varies from .19 to .26). With regard to inferential risks, the results show that in general the studies seem to have a good level of power (i.e., between 80% and 100%), except for the first and the last study which show lower power levels (i.e., 47% and 36%,

respectively). Moreover, these studies show also a risk of overestimating the plausible effect size of 43% and 59%, respectively. In fact, we can observe how the effect size of the first study is .50 and the effect size of the last study is .52, while the effect size of the meta-analysis on all studies excluded the one at hand (r_{meta}) is .22 and .23, respectively. It is interesting to note that these two studies are also the ones with the lowest sample sizes (i.e., 96 and 51, respectively) and the largest estimates of the effect size, which supports the claim that studies with low statistical power in order to obtain a statistically significant result need to estimate an effect size that might be “too big to be true” (Gelman & Carlin, 2014, see Chapter 1, section 1.1.3). Finally, these results underline the importance of conducting a retrospective design analysis and not judging a study based only on their statistical significance.

For statistical purpose only, we also used the “*Personalized*” method, assuming a plausible effect size of small dimension, i.e., $r = 0.1$, and a significant level of 0.5.

[illegible]

## 1	NA
## 2	NA
## 3	NA
## 4	NA
## 5	NA
## 6	NA
## 7	NA
## 8	NA
## 9	NA
## 10	NA
## 11	NA
## 12	NA

The results show that if the plausible effect size were $r = 0.1$., multiple studies would be underpower. For example, the first and the last show a remarkably low levels power, i.e., 16% and 11%, respectively. Moreover, compared to the analysis with the standard method, the M-type error also shows more variability. Specifically, it varies from low risks of overestimating the plausible effect size (i.e., 3% in the third study) to an expected overestimation almost three times and an half the plausible effect size (i.e., 236% in the last study).

5.3 Further Potentials Of The `drmeta` Function

`drmeta` function is a function created to explore the inferential risks associated with each study in a meta-analysis. However, there are several other potentials associated with this function.

Firstly, the variables that are obtained depending on the method performed (i.e., power, Type M and Type S errors) can be used as new variables to increase their

informativeness of results. For example, researchers can calculate descriptive statistics of the meta-analysis. In our case, the average power and the associated standard deviation (SD) of the meta-analysis conducted with the *standard corrected* method are power = 0.85 and SD = 0.22.

```
### loading the packages needed to run drmeta
library(PRDA)
library(metafor)
load("d.rda")
source("drmeta.R")

### recode data for drmeta
d1<- d[,-2]
names(d1)<-c("study","r","n")

### Assigning the results of power analysis of drmeta to an object
dg1<- drmeta(d1,method="stand_c",sig.level=.05)

### Calculating the average power and the associated SD
mean(dg1$power)

## [1] 0.8533333

sd(dg1$power)

## [1] 0.2172696
```

Secondly, the results of the meta-analysis conducted with the selected method can be graphically represented. For example, we chose to graphically represents the Type M errors associated with each study in the meta-analysis conducted with the “*Standard corrected*” method. To achieve this, we used a index plot (see figure 5.1 where the Type M error is reported in the x-axis and the identification code (ID) is reported in

the y-axis. The red dots represents the Type M error level for each study. As it can be seen, the first study and the last study have the highest risk of overestimating the plausible effect size (i.e., 43% and 59%, respectively).

Figure 5.1: The Type M errors for each study in the meta-analysis conducted with Standard corrected method

In the end, power, Type M and Type S errors can be used as new variables and associated with other variables on a study level. For example, referring to the original study, the author could be interested in exploring the association between the Type M error and the age of the sample or the year of publication.

Chapter 6

Conclusions

The low replication rate of research findings in psychology and the consequently credibility's crisis in the published results increased the awareness of the problematic issues in the literature. However, an important side-effect is that it promoted the development of new practices which would guarantee rigour, reproducibility, and transparency. For example, studies' pre-registration and pre-registered report, the creation of on-line long-term data repositories where to share materials and data (i.e., Open Science Frameworks) and the increased awareness that during the research process multiple factors must be taken into consideration (e.g., power, sample size, plausible effect size).

In the present work, the aim is to provide a small contribution analyzing the problem of studies with small sample sizes and small effects sizes (i.e., low power studies which are widespread in psychological literature). Specifically, we proposed a new statistical method, based on the power analysis approach developed by Gelman & Carlin (2014) called Design Analysis, to explore the inferential risks associated with each study in a meta-analysis.

The Design Analysis allows to evaluate two inferential risks, namely Type M and Type S errors and can be useful both during the research process and during results' evaluation. As discussed in the second chapter, the risks associated with underpowered

studies to detect a small effect size should not be underestimate: in order to detect a small effect size in an studies with a low power level, the real effect size must be overestimated and can have the wrong sign. In fact, the statistical significance could be the results of other factors rather than the presence of the true effect. This, combined with the presence of publication bias, which favors significant result over null or negative findings, could lead to misleading and unreliable results. These observation further highlight the gap between statistical significance and practical relevance on the results, which are the magnitude and the direction of the effects.

Moreover, during the research process is important to estimate what could be the plausible effect size of interest. Good practice should be to rely on different tools, such a extensive theoretical literature review, expert knowledge and/or on meta-analyses (Altoè et al., 2020), rather than on a single noisy study. However, meta-analysis could also report misleading results due to underpower studies or small sample size. These are the reason why we chose to proposed this new method, which can be conducted in R trough an ad doc function called *drmeta* that we developed. Researchers should be aware of the inferential risks associated with the data they are considering and consulting to make informed and conscious decisions.

However, although our function could be useful to analyse the inferential risks of each study in a meta-analysis, we are aware that there are other important factors that should be taken into consideration when dealing with meta-analysis (e.g, heterogeneity of studies).

It should also be noted that in the current work, we considered only one effect size measure (i.e., correlation coefficient) and one estimation measure, which is the random-effect meta-analysis with RML (i.e., Restricted Maximum Likelihood). However, these are limitations which can be easily overcome by future implementations of *drmeta*. In fact, the code (Appendix A) could be easily adapted to implement analysis also with other effect size and estimation measures.

In conclusion, this method can be an useful tool to increased the awareness of the potential risks associated with study results. Moreover, it highlights the relevance of reasoning about the plausible effect size and its crucial role before implementing the study as well as after results are known. In particular, reasoning about effect sizes could help to change researchers mindset from the dichotomous thinking focused on significant versus non-significant results to an estimation thinking based on the effect sizes and confidence intervals.

References

- Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagni, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology, 10*, 2893. <https://doi.org/10.3389/fpsyg.2019.02893>
- Bakker, M., Veldkamp, C. L., Akker, O. R. van den, Assen, M. A. van, Cromptvoets, E., Ong, H. H., & Wicherts, J. M. (2020). Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *Plos One, 15*(7), e0236079. <https://doi.org/10.1371/journal.pone.0236079>
- Becker, B. J. (2005). Failsafe N or file-drawer number. In *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–125). John Wiley.
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088–1101. <https://www.jstor.org/stable/2533446>
- Bertoldo, G. (2019). *Dealing with the replication crisis in psychological science: The contribution of Type M and Type S errors* [Preprint]. Thesis Commons. <https://doi.org/10.31237/osf.io/w63h7>
- Bertoldo, G., Zandonella Callegher, C., & Altoè, G. (2020). *Designing studies and evaluating research results: Type M and Type S Errors for pearson correlation coefficient* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/q9f86>

- Borenstein, M., & Hedges, H., L. V. (2009). *Introduction to meta-analysis*. John Wiley & Sons.
- Bushman, G. L., Brad J. and Wells. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin*, 27(9), 1123–1130. <https://journals.sagepub.com/doi/pdf/10.1177/0146167201279005>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.
- Callegher, C., Bertoldo, G., Toffalini, E., Vesely, A., Andreella, A., Pastore, M., & Altoè, G. (2021). PRDA: An R package for prospective and retrospective design analysis. *Journal of Open Source Software*, 6(58), 2810. <https://doi.org/10.21105/joss.02810>
- Chambers, C. D. (2019). *The sevens deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of “playing the game” it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(1), 4–17. <https://doi.org/10.3934/Neuroscience.2014.1.4>
- Cook, T. D., Campbell, D. T., & Shasish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Crocetti, E. (2016). Systematic reviews with meta-analysis: Why, when, and how? *Emerging Adulthood*, 4(1), 3–18. <https://doi.org/10.1177/2167696815617076>
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. <https://www.taylorfrancis.com/books/mono/10.4324/9780203807002/understanding-new-statistics-geoff-cumming>
- Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and repro-

- ducibility through registration: The status of intervention trials published in clinical psychology journals. *Journal of Consulting and Clinical Psychology*, 84(9), 753. [/doiLanding?doi=10.1037%2Fccp0000115](https://doi.org/10.1037%2Fccp0000115)
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.
- Duval, S. (2005). The trim and fill method. In *Publication bias in meta-analysis* (pp. 11–33). John Wiley.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109), 629–634. <https://www.bmj.com/content/315/7109/629.full>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://journals.sagepub.com/doi/full/10.1177/2515245920952393>
- Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, 105(2), 203–206. <https://doi.org/10.5195/JMLA.2017.88>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*. 17.
- Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, 5(10), 3–8. <https://doi.org/10.3102/0013189X005010003>
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps12–341ps12. <https://doi.org/10.1126/scitranslmed.aba3567>

[//doi.org/10.1126/scitranslmed.aaf5027](https://doi.org/10.1126/scitranslmed.aaf5027)

- 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>
- Haaf, J. M. (2020). *Conventional Publication Bias Correction Methods*. PsyArXiv. <https://doi.org/10.31234/osf.io/gv4tw>
- Heller, K., Price, R. H., Reinhartz, S., Riger, S., & Wandersman, A. (1989). *Psychology and community change: Challenges of the future, revised* (Second). Wadsworth Publishing.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Russell Sage Foundation.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Iso-Ahola, S. E. (2020). Replication and the establishment of scientific Truth. *Frontiers in Psychology*, 11, 2183. <https://www.frontiersin.org/article/10.3389/fpsyg.2020.02183/full>
- 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*, 2(3), 196–217. https://journals.sagepub.com/doi/abs/10.1207/s15327957pspr0203_4
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>
- Lakens, D. (2019). *The value of preregistration for psychological science: A conceptual analysis* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/jbh4w>
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*.

Harvard University Press.

- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585. <https://doi.org/10.1126/science.aal3618>
- Lu, J., Qiu, Y., & Deng, A. (2019). A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, 72(1), 1–17. <https://doi.org/10.1111/bmsp.12132>
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. <https://www.nap.edu/catalog/25303>
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, 18(3), e3000691. <https://dx.plos.org/10.1371/journal.pbio.3000691>
- Ohmer, M. L. (2007). Citizen participation in neighborhood organizations and its relationship to volunteers' self-and collective efficacy and sense of community. *Social Work Research*, 31(2), 109–120.
- Pashler, H., & Wagenmakers, E. (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://journals.sagepub.com/doi/full/10.1177/1745691612465253>
- Peng, R. D., Lionetti, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*, 163(9), 783–789. <https://academic.oup.com/aje/article/163/9/783/108733?login=true>
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638–641.

- Rother, E. T. (2007). *Systematic literature review X narrative review*.
- Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In *The handbook of research synthesis and meta-analysis* (2nd ed, pp. 103–125). Russell Sage Foundation.
- Santor, D. A., Gregus, M., & Welch, A. (2006). FOCUS ARTICLE: Eight decades of measurement in depression. *Measurement: Interdisciplinary Research and Perspectives*, 4(3), 135–155. https://www.tandfonline.com/doi/abs/10.1207/s15366359mea0403_1
- Schmidt, S. (2009). Shall we really do It again? The powerful concept of replication Is neglected in the social sciences. *Review of General Psychology*, 13, 90–100.
- Sedlmeier, P., & Gigerenzer, G. (1989). *Do studies of statistical power have an effect on the power of studies?* 8.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547. <https://doi.org/10.1037/a0033242>
- Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. In *Publication bias in meta-analysis* (pp. 75–98). John Wiley.
- Talò, C. (2018). Community-Based Determinants of Community Engagement: A Meta-Analysis Research. *Social Indicators Research*, 140(2), 571–596. <https://doi.org/10.1007/s11205-017-1778-y>
- Viechtbauer, W. (2010). Conducting Meta-Analyses in *r* with the **metafor** Package. *Journal of Statistical Software*, 36(3). <https://doi.org/10.18637/jss.v036.i03>
- Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical

CHAPTER 6. CONCLUSIONS

results. *PloS One*, 6(11), e26828. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026828>

Appendix A

drmeta Function

```
drmeta = function(d, method = c("stand",  
  "stand_c", "all", "all_c", "pers"), r = NA,  
  B = 10000, sig.level = 0.05) {  
  #  
  library(metafor)  
  library(PRDA)  
  #  
  method = match.arg(method)  
  if ((method == "pers") & is.na(r))  
    stop("If method = pers, you must specify the hypothesized r  
    value!")  
  d = data.frame(id = c(1:nrow(d)), d)  
  # Fisher transformation  
  dt = d  
  dt$r.zi = 0.5 * log((1 + dt$r)/(1 - dt$r))  
  dt$r.vi = (1/(dt$n - 3)) # OK  
  dt$power = NA
```

APPENDIX A. DRMETA FUNCTION

```
dt$type_M = NA
dt$type_S = NA
dt$r_meta = NA
dt$r_meta_cor = NA

#### Implementation of the 5
#### methods Standard method
if (method == "stand") {
  fisher.z = rep(NA, nrow(dt))
  for (i in 1:nrow(dt)) fisher.z[i] = as.numeric(rma(dt[-i,
    ]$r.zi, dt[-i, ]$r.vi, method = "REML")$b)
  dt$r_meta = tanh(fisher.z)
  #
  for (i in 1:nrow(dt)) {
    temp = retrospective(effect_size = dt$r_meta[i],
      sample_n1 = dt$n[i], test_method = "pearson",
      B = B, sig_level = sig.level)$retrospective_res
    dt[i, c(7:9)] = round(as.numeric(temp),
      2)
  }
  return(dt[, -c(5, 6)])
}

#### Standard corrected method
if (method == "stand_c") {
  fisher.z = rep(NA, nrow(dt))
  fisher.z_c = fisher.z
  for (i in 1:nrow(dt)) {
```

APPENDIX A. DRMETA FUNCTION

```
fisher.z[i] = as.numeric(rma(dt[-i,
                             ]$r.zi, dt[-i, ]$r.vi, method = "REML")$b)
fisher.z_c[i] = as.numeric(trimfill(rma(dt[-i,
                             ]$r.zi, dt[-i, ]$r.vi, method = "REML"))$b)
}

dt$r_meta = tanh(fisher.z)
dt$r_meta_cor = tanh(fisher.z_c)

for (i in 1:nrow(dt)) {
  temp = retrospective(effect_size = dt$r_meta_cor[i],
                      sample_n1 = dt$n[i], test_method = "pearson",
                      B = B, sig_level = sig.level)$retrospective_res
  dt[i, c(7:9)] = round(as.numeric(temp),
                        2)
}

return(dt[, -c(5, 6)])
}

#### All studies method
if (method == "all") {
  fisher.z = as.numeric(rma(dt$r.zi,
                          dt$r.vi, method = "REML")$b)
  r = tanh(fisher.z)
  dt$r_meta = r
  for (i in 1:nrow(dt)) {
    temp = retrospective(effect_size = r,
                        sample_n1 = dt$n[i], test_method = "pearson",
```


APPENDIX A. DRMETA FUNCTION

```
        B = B, sig_level = sig.level)$retrospective_res
    dt[i, c(7:9)] = round(as.numeric(temp),
        2)
  }
  return(dt[, -c(5, 6)])
}

#### All studies corrected method
if (method == "all_c") {
  fisher.z = as.numeric(rma(dt$r.zi,
    dt$r.vi, method = "REML")$b)
  fisher.z_c = as.numeric(trimfill(rma(dt$r.zi,
    dt$r.vi, method = "REML"))$b)
  dt$r_meta = tanh(fisher.z)
  dt$r_meta_cor = tanh(fisher.z_c)
  for (i in 1:nrow(dt)) {
    temp = retrospective(effect_size = dt$r_meta_cor[1],
      sample_n1 = dt$n[i], test_method = "pearson",
      B = B, sig_level = sig.level)$retrospective_res
    dt[i, c(7:9)] = round(as.numeric(temp),
      2)
  }
  return(dt[, -c(5, 6)])
}

#### Personalized
if (method == "pers") {
  dt$r_meta = r # NB
```

APPENDIX A. DRMETA FUNCTION

```
for (i in 1:nrow(dt)) {  
  temp = retrospective(effect_size = r,  
    sample_n1 = dt$n[i], test_method = "pearson",  
    B = B, sig_level = sig.level)$retrospective_res  
  dt[i, c(7:9)] = round(as.numeric(temp),  
    2)  
}  
return(dt[, -c(5, 6)])  
}  
####  
}
```

Appendix B

How drmeta function works

Here is described the R function defined to explore the inferential risks (i.e., power, Type M and Type S errors) associated with each study in a meta-analysis: `drmeta`. Initially, instructions on how to load and use the function are given. Then, the code to reproduce examples included in the article will be provided.

To use `drmeta` function, first download the file available in Appendix A. To load the function, type:

```
source("drmeta.R")
```

Note: For the correct use of the function, the R packages `metafor` (Viechtbauer, 2010) and `PRDA` (Callegher et al., 2021) must be installed.

The function is:

```
drmeta(d, method = c("stand", "stand_c", "all", "all_c",  
  "pers"), r = NA, B = 10000, sig.level = 0.05)
```

The function `drmeta()` runs inferential risks (i.e., power, Type M and Type S errors) associated with each study in a meta-analysis according to a fixed type I error (`sig.level`). Specifically, users must specified:

APPENDIX B. HOW DRMETA FUNCTION WORKS

- data dataframe must consist of three columns showing study, coefficient correlation (r) and sample size (n), in exactly this order

Note. Variable names must be “study,” “r” and “n.”

- method user can choose between 5 methods “stand,” “stand_c,” “all,” “all_c,” “pers”

Note. If users decide to use the “pers” method, they must specify a coefficient correlations (r).

Function arguments

- d data
- method a character string specifying the method user can choose to run, must be one of “stand,” “stand_c,” “all,” all_c” or “pers.” Default is “stand”
- r coefficient correlation. Default is NA.
- B number of replications for simulations. Default is 10000
- sig.level type I error. Default is 0.5

Function returned values - A list containing main arguments and results

Examples code

Below the code to reproduce the examples presented in chapter 4 are reported:

```
source("drmeta.R")
study=c("Study1","Study2","Study3","Study4","Study5","Study6")
r= c(.80,.30,.30,.65,.70,.10)
n = c(10,45,100,60,30,20)
dati=data.frame(study,r,n)
dati
```

```
##      study      r      n
```

APPENDIX B. HOW DRMETA FUNCTION WORKS

```
## 1 Study1 0.80 10
## 2 Study2 0.30 45
## 3 Study3 0.30 100
## 4 Study4 0.65 60
## 5 Study5 0.70 30
## 6 Study6 0.10 20
```

```
source("drmeta.R")
drmeta(d=dati,method="stand", sig.level = 0.5)
```

##	id	study	r	n	power	type_M	type_S	r_meta	r_meta_cor
## 1	1	Study1	0.80	10	0.79	1.22	0.03	0.4478691	NA
## 2	2	Study2	0.30	45	1.00	0.99	0.00	0.5307045	NA
## 3	3	Study3	0.30	100	1.00	1.00	0.00	0.5365004	NA
## 4	4	Study4	0.65	60	1.00	1.00	0.00	0.4428927	NA
## 5	5	Study5	0.70	30	0.96	1.01	0.00	0.4353529	NA
## 6	6	Study6	0.10	20	0.97	1.01	0.00	0.5392838	NA

##	id	study	r	n	power	type_M	type_S	r_meta	r_meta_cor
## 1	1	Study1	0.80	10	0.26	1.65	0	0.4478691	0.4478691
## 2	2	Study2	0.30	45	0.97	1.01	0	0.5307045	0.5307045
## 3	3	Study3	0.30	100	1.00	1.00	0	0.5365004	0.5365004
## 4	4	Study4	0.65	60	0.86	1.07	0	0.4428927	0.3790303
## 5	5	Study5	0.70	30	0.69	1.17	0	0.4353529	0.4353529
## 6	6	Study6	0.10	20	0.72	1.14	0	0.5392838	0.5392838

##	id	study	r	n	power	type_M	type_S	r_meta	r_meta_cor
## 1	1	Study1	0.80	10	0.31	1.52	0	0.4892741	NA
## 2	2	Study2	0.30	45	0.94	1.03	0	0.4892741	NA

APPENDIX B. HOW DRMETA FUNCTION WORKS

```
## 3 3 Study3 0.30 100 1.00 0.99      0 0.4892741      NA
## 4 4 Study4 0.65 60 0.98 1.00      0 0.4892741      NA
## 5 5 Study5 0.70 30 0.80 1.09      0 0.4892741      NA
## 6 6 Study6 0.10 20 0.62 1.21      0 0.4892741      NA
```

```
## id study r n power type_M type_S r_meta r_meta_cor
## 1 1 Study1 0.80 10 0.26 1.65      0 0.4892741 0.4473219
## 2 2 Study2 0.30 45 0.88 1.06      0 0.4892741 0.4473219
## 3 3 Study3 0.30 100 1.00 1.00      0 0.4892741 0.4473219
## 4 4 Study4 0.65 60 0.96 1.02      0 0.4892741 0.4473219
## 5 5 Study5 0.70 30 0.72 1.15      0 0.4892741 0.4473219
## 6 6 Study6 0.10 20 0.53 1.30      0 0.4892741 0.4473219
```

```
## id study r n power type_M type_S r_meta r_meta_cor
## 1 1 Study1 0.80 10 0.11 2.89 0.05 0.25      NA
## 2 2 Study2 0.30 45 0.39 1.55 0.00 0.25      NA
## 3 3 Study3 0.30 100 0.71 1.18 0.00 0.25      NA
## 4 4 Study4 0.65 60 0.50 1.39 0.00 0.25      NA
## 5 5 Study5 0.70 30 0.27 1.83 0.00 0.25      NA
## 6 6 Study6 0.10 20 0.18 2.16 0.01 0.25      NA
```

Appendix C

The Case study

```
library(readxl)
d <- read_excel("Table2.xlsx", sheet = "recode")
d$cor<-as.numeric(d$cor)
d<-data.frame(d)
d<-d[,-1]
d
```

```
##              authors year   cor    n
## 1              Chang 2010 0.500   96
## 2    Chioneso and Brookins 2015 0.290  235
## 3    Christens and Lin 2014 0.090 1322
## 4    Cicognani et al. (IR) 2008 0.202  214
## 5    Cicognani et al. (IT) 2008 0.261  200
## 6    Cicognani et al. (USA) 2008 0.231  125
## 7    Mannarini et al. 2010 0.190  194
## 8    Mazzoni et al. 2014 0.245  835
## 9    Moreno-Jiménez et al. 2013 0.326  756
```

APPENDIX C. THE CASE STUDY

```
## 10 Ramos-Vidal and Maya Jariego 2014 0.010 120
## 11          Serek and Machackova 2015 0.190 573
## 12          Tsai et al 2008 0.516 51
```

```
save(d,file="d.rda")
```

```
library(PRDA)
library(metafor)
library(psych)
load("d.rda")
source("drmeta.R")
d
```

```
##          authors year  cor   n
## 1          Chang 2010 0.500  96
## 2    Chioneso and Brookins 2015 0.290 235
## 3    Christens and Lin 2014 0.090 1322
## 4    Cicognani et al. (IR) 2008 0.202 214
## 5    Cicognani et al. (IT) 2008 0.261 200
## 6    Cicognani et al. (USA) 2008 0.231 125
## 7    Mannarini et al. 2010 0.190 194
## 8    Mazzoni et al. 2014 0.245 835
## 9    Moreno-Jiménez et al. 2013 0.326 756
## 10 Ramos-Vidal and Maya Jariego 2014 0.010 120
## 11    Serek and Machackova 2015 0.190 573
## 12    Tsai et al 2008 0.516 51
```


APPENDIX C. THE CASE STUDY

```
#### recode data for drmeta
```

```
d1<- d[,-2]
```

```
names(d1)<-c("study","r","n")
```

```
#
```

```
#####
```

```
r.zi= .5 * log ( (1+d1$r) / (1-d1$r) )
```

```
r.vi= (1 / (d1$n-3))
```

```
model1<-rma(yi=r.zi,vi=r.vi,method ="REML")
```

```
round( fisherz2r(model1$b) , 3)
```

```
##           [,1]
```

```
## intrcpt 0.243
```

```
round( fisherz2r(c(model1$ci.lb , model1$ci.ub)) , 3)
```

```
## [1] 0.173 0.311
```

```
trimfill(model1)
```

```
##
```

```
## Estimated number of missing studies on the left side: 0 (SE = 2.2480)
```

```
##
```

```
## Random-Effects Model (k = 12; tau^2 estimator: REML)
```

```
##
```

```
## tau^2 (estimated amount of total heterogeneity): 0.0120 (SE = 0.0070)
```

```
## tau (square root of estimated tau^2 value):      0.1096
```

```
## I^2 (total heterogeneity / total variability):   81.09%
```

```
## H^2 (total variability / sampling variability):   5.29
```

```
##
```

APPENDIX C. THE CASE STUDY

```
## Test for Heterogeneity:
## Q(df = 11) = 57.0118, p-val < .0001
##
## Model Results:
##
## estimate      se      zval      pval      ci.lb      ci.ub
## 0.2477 0.0375 6.6094 <.0001 0.1743 0.3212 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####
```

```
drmeta(d1,method="stand_c",sig.level=.05)
```

```
##      id              study      r      n power type_M type_S      r_meta
## 1  1              Chang 0.500    96  0.47   1.43        0 0.2219230
## 2  2      Chioneso and Brookins 0.290  235  0.96   1.02        0 0.2395236
## 3  3      Christens and Lin 0.090 1322  1.00   1.00        0 0.2590618
## 4  4      Cicognani et al. (IR) 0.202  214  0.96   1.02        0 0.2479902
## 5  5      Cicognani et al. (IT) 0.261  200  0.94   1.03        0 0.2426163
## 6  6      Cicognani et al. (USA) 0.231  125  0.79   1.12        0 0.2450239
## 7  7      Mannarini et al. 0.190  194  0.94   1.03        0 0.2488134
## 8  8      Mazzoni et al. 0.245  835  1.00   1.00        0 0.2444694
## 9  9      Moreno-Jiménez et al. 0.326  756  1.00   1.00        0 0.2330830
## 10 10 Ramos-Vidal and Maya Jariego 0.010  120  0.82   1.11        0 0.2565792
## 11 11      Serek and Machackova 0.190  573  1.00   1.00        0 0.2501538
## 12 12      Tsai et al 0.516    51  0.38   1.59        0 0.2283708
##      r_meta_cor
```

APPENDIX C. THE CASE STUDY

```
## 1 0.1918136
## 2 0.2395236
## 3 0.2590618
## 4 0.2479902
## 5 0.2426163
## 6 0.2450239
## 7 0.2488134
## 8 0.2444694
## 9 0.2330830
## 10 0.2565792
## 11 0.2501538
## 12 0.2283708
```

```
drmeta(d1,method="pers",r=.1,sig.level=.05)
```

##	id	study	r	n	power	type_M	type_S	r_meta
## 1	1	Chang	0.500	96	0.16	2.51	0.01	0.1
## 2	2	Chioneso and Brookins	0.290	235	0.33	1.71	0.00	0.1
## 3	3	Christens and Lin	0.090	1322	0.96	1.03	0.00	0.1
## 4	4	Cicognani et al. (IR)	0.202	214	0.31	1.77	0.00	0.1
## 5	5	Cicognani et al. (IT)	0.261	200	0.29	1.82	0.00	0.1
## 6	6	Cicognani et al. (USA)	0.231	125	0.19	2.22	0.00	0.1
## 7	7	Mannarini et al.	0.190	194	0.28	1.84	0.00	0.1
## 8	8	Mazzoni et al.	0.245	835	0.82	1.11	0.00	0.1
## 9	9	Moreno-Jiménez et al.	0.326	756	0.79	1.13	0.00	0.1
## 10	10	Ramos-Vidal and Maya Jariego	0.010	120	0.20	2.27	0.01	0.1
## 11	11	Serek and Machackova	0.190	573	0.67	1.22	0.00	0.1
## 12	12	Tsai et al	0.516	51	0.11	3.36	0.04	0.1

APPENDIX C. THE CASE STUDY

```
##      r_meta_cor
## 1          NA
## 2          NA
## 3          NA
## 4          NA
## 5          NA
## 6          NA
## 7          NA
## 8          NA
## 9          NA
## 10         NA
## 11         NA
## 12         NA
```

```
##
dg1<- drmeta(d1,method="stand_c",sig.level=.05)
# descriptive statistics
mean(dg1$power)
```

```
## [1] 0.8533333
```

```
sd(dg1$power)
```

```
## [1] 0.2153785
```

```
# graphs type M
plot(dg1$id,dg1$type_M,axes=F,ylab="type M error",xlab="index",
      pch=16,col="red")
axis(2)
```

APPENDIX C. THE CASE STUDY

```
axis(1,at=c(1:nrow(dg1),labels=dg1$id))  
abline()  
box()
```