UNIVERSITY OF PADOVA

Department of Developmental Psychology and Socialisation

Master Degree in
Developmental and Educational Psychology

Final Dissertation

Multiverse Meta-Analysis: Proposing an Exploratory Framework

Supervisor:
Professor Gianmarco Altoè

Co-supervisor:
Dr. Filippo Gambarota

Candidate: Sajedeh Rasti
Student ID number: 2042056

Academic year 2022/2023

Acknowledgement

I want to express my heartfelt gratitude to Professor Altoè for his exceptional support, unwavering guidance, and immense generosity throughout my journey. His support has extended far beyond the boundaries of traditional mentorship. His insightful guidance has not only broadened my knowledge but has also challenged me to push beyond my limits and strive for excellence. His willingness to invest time and effort in my progress has demonstrated his genuine commitment to my success.

I am truly indebted to Professor Lakens. His constant encouragement and unwavering faith in my abilities have ignited and fueled my passion for this field of research. During times of confusion and uncertainty, Professor Lakens has been a guiding light, providing invaluable guidance and direction.

I am eternally grateful to my parents, whose love and acceptance have shaped me into the person I am today. From their unshakeable resilience and determination to their unyielding courage and hard work, I have learned invaluable lessons. Their support has provided me with the freedom to explore, grow, and flourish. Their constant belief in me has empowered me to seize life-changing opportunities and embrace my unique qualities. I am forever indebted for all these gifts.

To my second family, my dear friends, I cannot thank you enough for being my pillars of strength during my darkest hours. You have played an indispensable role in my personal growth, empowering me to overcome obstacles and become the best version of myself. Thank you, from the bottom of my heart, for being such amazing friends.

To my dear Alireza, you have been a source of inspiration, constantly pushing me forward to reach new heights. It’s difficult to recall a time when you weren’t a part of my life. I am immensely grateful for all the invaluable lessons you have taught me throughout the years. Your presence has had a transformative impact on my journey, and I cherish the moments we have shared. Thank you for being an incredible influence.

Table of Contents

Summary

Recognition of the credibility crisis shed light on proposals to alleviate its impact. One of these proposals was multiverse analysis (Steegen et al., 2016). The main idea of this analysis is to consider all plausible choices in the flow of research to assess the robustness and informativeness of the results. Implementing this idea is particularly beneficial in meta-analytical studies. Not only do these studies usually have a high impact on literature, but they also involve making multiple decisions that are usually taken arbitrarily. However, many researchers still struggle to handle and communicate numerous results of this approach and consider multiverse results to be challenging and overwhelming.

The present study aims to address this challenge by introducing an exploratory framework to assist researchers in evaluating and communicating the result of multiverse meta-analysis using tabular and graphical representations. Furthermore, we will highlight the contribution of each arbitrary decision to the variability of results.

The first chapter will discuss the role of science as well as the consequences and reasons for the current credibility crisis. It also covers some of the remedies to alleviate the impact of this crisis. The second chapter will focus on the multiverse analysis. It will introduce this analysis along with similar proposals, and it will introduce other contexts that can benefit from the multiverse approach. Chapter three will be dedicated to introducing the meta-analysis with a focus on arbitrary choices that are involved in the flow of the meta-analytical research. In particular, we will discuss choices concerning

the effect size, meta-analytical model, heterogeneity, and bias control. In Chapter Four, we will introduce our exploratory framework, which will be further explained by implementing it on a real dataset in Chapter Five. Finally, we will conclude by recognizing the benefits and limitations of this framework in Chapter Six.

This thesis was written in R markdown.

Chapter 1

Credibility Crisis

What is the role of science? We should be clear about what is expected of science in general before starting to tackle the problems circulating scientific research. According to Gibbons (1999), the role of science is to transparently produce reliable knowledge. If we accept this statement as a general role of science, it is evident that trust plays an important part in the fulfillment of this role. Despite the notion that public trust in science remained stable over time (Scheufele, 2013), science has been facing challenges as a result of overlooking this trust (Rutjens et al., 2018).

Credibility crisis is a general term referring to a lack of confidence in results drawn by research, especially in the fields of social and biomedical sciences (Gall et al., 2017). This distrust is fueled by many evidence of the lack of replicability of results obtained by many research studies (Bettis, 2012; Ioannidis, 2012; Open Science Collaboration, 2015). If continued, the consequences of this increasing mistrust can be unbearable. It can result in endangering people's well-being, increasing anti-science movements (e.g., Ioannidis, 2017), and retracting funds. As no one can become an expert in all fields of science (Anvari & Lakens, 2018), protecting trust should be the primary responsibility of everyone working in this field.

As also mentioned before, replicability plays an important role in building confidence and credibility for science (Hendriks et al., 2020; National Academies of Sciences et al., 2019). However, it is vital to distinguish the difference between replicability and reproducibility in science first. These two terms refer to fundamental characteristics of science (Patil et al., 2016). Unfortunately, many scientists use these words interchangeably, yet, it is useful and essential to know the difference. *Replicability* refers to " re-performing the experiment and collecting new data", whereas *reproducibility* refers to " re-performing the same analysis with the same code using a different analyst" (Patil et al., 2016). Given this definition, reproducibility is directly connected to transparency, and it needs the original author to transparently report the entire process of analysis including data, code, and analysis plan (National Academies of Sciences et al., 2019; Stevens, 2017). Replicability is also related to transparency, but the extent to which it is related varied according to different kinds of replicability. The logic of *direct replicability* is to repeat the method of the original study as precisely as possible to get consistent estimates of the original study (Chambers, 2017). As it is evident, this kind of replication requires a transparent and adequate report of statistical and methodological details (Derksen & Morawski, 2022). On the other hand, *conceptual replicability* refers to replicating the theory of the original study using different comparable operationalizations, variables, and experimental designs (Derksen & Morawski, 2022).

It is believed that conceptual replication is specifically important in the field of psychology, as many psychological phenomena are sensitive to context and as a result, cannot be replicated directly (Derksen & Morawski, 2022). However, according to Chambers (2017) relying on conceptual replication is not without challenges. Firstly, the extent to which the methods are comparable in the original and the replicated study is a matter of controversy. Secondly, whether the results reported by both studies are in fact supporting the same phenomenon is not always clear; And lastly, with conceptual replication, there is always room for confirmation bias, as the original author can always blame different methods for the different conclusion drawn by the replication study (Chambers, 2017).

In this chapter, we will review some of the potential reasons for mistrust and credibility crisis and discuss some possible remedies to retrieve the credibility of science. In the end, we will conclude with an overview of the goals of the present study.

1.1 Possible reasons for credibility crisis

Several flawed practices attribute to ruining the trust and credibility of science. Not having comprehensive knowledge about these threats will result in continual repetition of these actions and undermining of the scientific conclusions. These practices include bending the data to produce publishable outcomes, misusing statistical tools and failing to bring justification when necessary, and using invalid measurements. In this section, we will discuss each of these contributing factors separately.

1.1.1 Questionable research practices

For more than 60 years, the scientific community has been aware of questionable research practices (QRPs) (Banks et al., 2016). Unfortunately for science, only in the last three years, the prevalence of these practices was as high as 17.5%, with more than 50% of participant researchers engaging frequently in at least one QRPs (Gopalakrishna et al., 2022). According to Banks et al. (2016), questionable research practice is defined as “design, analytic, or reporting practices that have been questioned because of the potential for the practice to be employed with the purpose of presenting biased evidence in favor of an assertion” (p. 3). The incentive of engaging in QRPs for many scientists is to increase the chance of their papers getting published (Chambers, 2017).

Not only using these practices is unethical and misleading, but also they have a damaging influence on public trust, as they contribute to the irreplicability of science

(Gopalakrishna et al., 2022). Questionable research practices can adopt different forms: lack of transparency, presenting incomplete evidence, and intentional misrepresentation of data (Flake & Fried, 2020). Considering the hypothetico-deductive model, QRMs can happen during the collecting, analyzing, and interpreting of data. In any case, in the end, they contribute to research bias (Chambers, 2017).

P-hacking or “selective reporting” is one of the most famous questionable research practices. It is defined as manipulating data and/or statistical analysis to include only those resulting in statistically significant outcomes (Head et al., 2015; Raj et al., 2017). As it is apparent, p-hacking has a close relationship with the null hypothesis significant testing (NHST) (Chambers, 2017). According to NHST, the p-value is the probability of observing the obtained data or more extreme data if the null hypothesis is true (Lakens, 2021). In order to make inferential decisions, the obtained p-value is compared with a predefined α value, to either reject or not reject the null hypothesis (Chambers, 2017). Unfortunately, as many researchers, as well as journals, have a predilection for statistically significant results, many get involved in p-hacking to obtain significant results. The prevalence of p-hacking in some fields, like psychology is so high that it can be seen as a norm (Chambers, 2017).

P-hacking has several forms: optional stopping, choosing to include specific data points among many obtained data, including or excluding outliers and/or covariates post-analyses, changing in the treatment groups post-analyses, and stopping data exploration after yielding statistically significant results (Head et al., 2015).

When any of the aforementioned forms of p-hacking found their way into the published literature, the impact will be substantial and persistent. These misconducted papers will contribute to overestimating the true effect size, as well as inspiring other fruitless research programs and spoiling public funds (Head et al., 2015). Regrettably, even when replicating the study (if there would be any) provides proof for misleading results, the replication study would never get as enough attention as the previous one (Chambers, 2017; Head et al., 2015).

HARKing (Hypothesizing After Results are Known) is another widespread QRP. This form of academic deception refers to when the author changes the initial hypothesis after analyzing data (Kerr, 1998). This changes include but are not limited to reversing the direction of the hypothesis(Chambers, 2017), retrieving the previously proposed hypothesis, and not presenting the non-favorable a priori hypothesis which is supported by the data (Lishner, 2021). As publishing the hypothesis is rarely done in advance, altering hypothesis can be done without being noticed (Chambers, 2017). However, some evidence can trace this action; among those are too convenient qualifiers, too-good-to-be-true theory, and poorly fitted design (see Kerr, 1998). In any case, the prevalence of HARKing can go as high as 90% although the self-admission rate is much lower (Chambers, 2017; Murphy & Aguinis, 2019).

HARKing has negative impacts on different aspects of science. First of all, it betrays the ethical principle of honesty in science and risks the trustworthiness of science by an extension (Chambers, 2017). From a theoretical perspective, HARKing can spread the trend of unfalsifiable and/or unnecessary complex theoretical explanations (Lishner, 2021). Additionally, it misleads the researchers to over-rely on nonindependent explanations (Chambers, 2017; Lishner, 2021). Lastly, by increasing inaccurate and mismatched illustrations of the scientific process, HARKing will jeopardize the integrity of scientific methodology (Lishner, 2021).

The reasons behind HARKing can be classified into two main clusters. Firstly, many scientific authors have a biased predisposition toward the scientific approach. Not only do they account for less value in exploratory research, they nearly always find confirmation as more valuable and instructive (Kerr, 1998). Secondly, HARKed research reports are contributing to a hypothetical picture of a “good story” in science. They provide an illusion that all activities and tests done by the researcher were purposeful and justified and would come to a “happy ending” (Kerr, 1998).

The last QRP we will describe is *cherry-picking*. In the field of science, cherry-picking refers to the selective report of findings that have the strongest possible support for the hypothesis (Murphy & Aguinis, 2019). In the most common form of this practice, researchers go through the data and pick only significant outcomes and overlook insignificant ones as if they had not been studied (Andrade, 2021). This usually happens because the researcher completes the analysis with an inadequate number of data (Morse, 2010). Another form of cherry-picking happens when the researcher only considers and cites those papers that favor their argument (Andrade, 2021). In both cases, cherry-picking represents the selective attention of the author which is drawn by confirmation bias (Elston, 2021).

As mentioned before, cherry-picking is a way to find support for the researchers' prior beliefs. Of course, when authors make a mistake in their prior beliefs, faulty results will find their way into the literature (Morse, 2010). According to Murphy & Aguinis (2019), cherry-picking will always produce biased results, however, they believed it would not generally contribute to publication bias. On the other hand, Mayo-Wilson et al. (2017) reported that the bias made by this practice would substantially interfere with the significance and magnitude of meta-analytic results (Mayo-Wilson et al., 2017). On the whole, cherry-picking can misguide many other scientists in their further research and interventions.

The domain of questionable research practices is vast. We did not cover data fabrication in this section as its damaging influence is evident even for the general public. Any intentional or accidental involvement in such practices will attribute to the credibility crisis and violation of the code of conduct in research practices. Comprehensive knowledge of different forms and the impact of such practices will reduce engagement in them and hopefully, would mitigate their effects on scientific literature.

1.1.2 Questionable research design

The proposed name of “questionable research design” refers to overlooking important details during the research design that can contribute to irreplicability and unfalsifiability. Long before making inferences about the data and engaging in any form of questionable research practices, researchers make decisions, or in some cases do not make certain necessary decisions, about the design and analysis of their study. These decisions not only fuels the transparency problems in research practice but reduce the dignity and reliability of their analysis. In this section, we will review some of these practices.

Although the Neyman-Pearson approach to hypothesis testing is one of the most dominant approaches in science, many researchers seem to have misunderstood the fundamental details of this approach. Among all aspects of NHST, probably the most misunderstood one is the concept of the p-value (Chambers, 2017). However, a step before misinterpretation of p-values is *deciding on α and β values.* We might assume that when most researchers choose the universal values of $\alpha = 0.05$ and $\beta = 0.20$, we can rely on their decision, which in the eyes of a mature scientist, is not the case (Maier & Lakens, 2022). The idea of universal α value neither was supported by Fisher nor Neyman and Pearson (Maier & Lakens, 2022). On the contrary, they emphasized the researchers' role in minimizing and controlling the risk of error (Maier & Lakens, 2022). However, a misunderstanding from Fisher's note led to the conventional value of 0.05 for the alpha value (see Maier & Lakens, 2022). Later, based on this conventional value for α, Cohen (1988) proposed his preferable value for beta, $\beta = 0.20$, only **“when the investigator has no other basis for setting the desired power value”** (p. 56). The universal use of these conventions has been criticized for a long time (Maier & Lakens, 2022). The problems with these universal values are that they will reduce the efficiency of our decision making and they can increase the probability of committing an error in some cases (Maier & Lakens, 2022). To tackle these problems, Benjamin et al. (2018) proposed to reduce the alpha level to $\alpha = 0.005$ in descriptive research in certain fields, so by instantly lowering the probability of committing a type 1 error, the replicability will improve. As this way would increase the probability of a false negative, Benjamin et al. (2018) suggested increasing the sample size to keep the power constant. However, this solution is not without critics. Lakens, Adolfi, et al. (2018) argued that by putting another universal constant value for alpha, we are not solving the problem. As long as p-hacking is confounding the result of the studies, we will not solve the problem of replicability only by lowering the alpha level (Lakens, Adolfi, et al., 2018). However, we should note that p-hacking would be a much easier job when the alpha level is 0.05 compared to the proposed level (Ruiter, 2019).

Another way to tackle the aforementioned problems is to **justify** alpha and beta values (Lakens, Adolfi, et al., 2018). Two practical ways to do so are to balance or minimize the cost of type 1 and type 2 error rates, and the second is to lower the alpha level as a function of sample size (Maier & Lakens, 2022). In cases where none of the error types are significantly more costly, we can design the study in a way to have the minimum Weighted Combined Error rate (Maier & Lakens, 2022). Through this approach, by considering essential factors of hypothesis testing, such as sample size and prior belief, we can choose a more optimal level of significance (Kim & Choi, 2021). There is only one alpha level that can minimize the weighted combined error rate given a specific sample size and effect size (Figure 1.1). Therefore, by sticking to $\alpha = 0.05$, we only increase the probability of committing an error.

The second practical way to justify the alpha level is mostly used to tackle the problem of **Lindley's Paradox**. According to Lindley's paradox, when we have high sample sizes (a.k.a. high statistical power) the probability of observing some p-values below 0.05 is higher when the null hypothesis is true (Maier & Lakens, 2022). The idea of this approach is to decrease the alpha value in a way that the Bayes factor $(\frac{p(data|H_1)}{p(data|H_0)})$

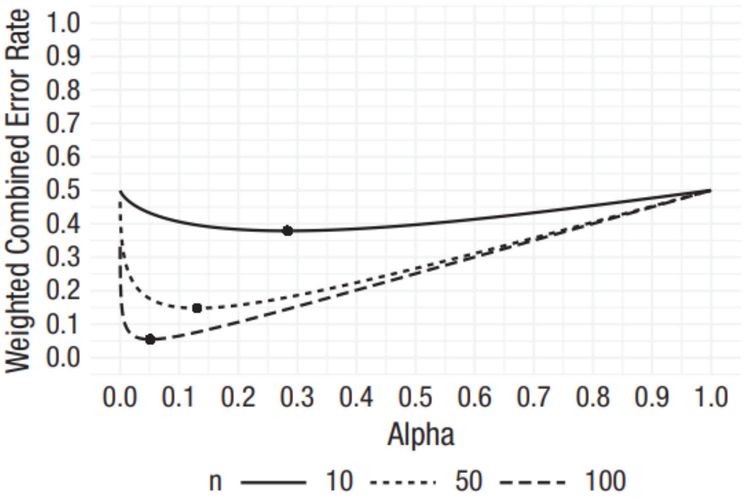


Figure 1.1: Weighted combined error rate for an independent t test with different sample sizes for all possible alpha values. Adapted from Justify your alpha: A primer on two practical approaches by M. Maier and D. Lakens, 2022, Advances in Methods and Practices in Psychological Science, 5 (2), P. 4

would be more than 1 (e.g., BF>3); by doing so, we would have stronger evidence that the significant data is not a case of Lindley's paradox (see Maier & Lakens, 2022). Although most aforementioned approaches were more focused on changing the alpha value and Bartoš & Maier (2022) also argue it is more efficient to change the alpha level to change the statistical power, we cannot overlook the problems caused by *low statistical power*. Many researchers have already uncovered the low statistical power of psychological findings (Bartlett & Charles, 2022; Chambers, 2017). This means that studies are searching for smaller effect sizes than the ability of their design (Bartlett & Charles, 2022). In other words, by underpowered studies, true effects cannot be found often (Nuijten, 2019), and as it is not the case in the literature, we can see how low-powered studies can affect the reliability of science (Chambers, 2017). Furthermore, this lack of sensitivity to detect the true effect can delay theory generation for a long time and make researchers wrongly abandon the correct hypothesis (Chambers, 2017). One way to mitigate this problem is to recruit enough participants for the study; this process is called "power analysis" (Bartlett & Charles, 2022). This approach is not the only way to justify the sample size (see Lakens, 2022), but it is a way to avoid low statistical power.

Another practice that is threatening the credibility of results drawn by research is the failure to determine the *Smallest Effect Size of Interest* (SESOI). Statistical significance does not give us any information about the practicality of the result, yet we can draw such information from the effect size (Greenland et al., 2016). Using effect sizes has become an important indicator of planning and interpreting studies, especially in psychological science (Riesthuis et al., 2022). However, the tools to interpret the meaningfulness of the effect sizes seem to be scarce (Anvari & Lakens, 2021). One of these tools is to compare the obtained result to the SESOI. There are three ways to decide on the smallest effect size of interest: a) theoretical relevance, b) cost-benefit analysis, and c) the researcher's personal choice (Lakens, 2014; Lakens, Scheel, et al., 2018). In any case, the value of SESOI is independent of the research result and should be decided before looking at the data (Lakens, Scheel, et al., 2018). Determining SESOI can improve the credibility of research to a reasonable extent. The most important benefit of SESOI in my view is that it provides the possibility to design falsifiable studies (Anvari & Lakens, 2021), which is a fundamental characteristic of replicability. Moreover, it can help researchers to choose sample size in a way to have high power to detect meaningful effects (Anvari & Lakens, 2021; Lakens, 2022). Additionally, it will help researchers to decide the absence of large enough (meaningful) effects by using Equivalence testing (Anvari & Lakens, 2021), and as a result, prevent further spending funds on dead-end topics.

Like QRPs, the different forms of questionable research designs are broad. As it was evident from this section, many of these design choices are interconnected and will influence one another. Scientific research is not a roll of dice; it is essential to design the study in a way that would be the most informative, regardless of the result.

1.1.3 Validation Crisis

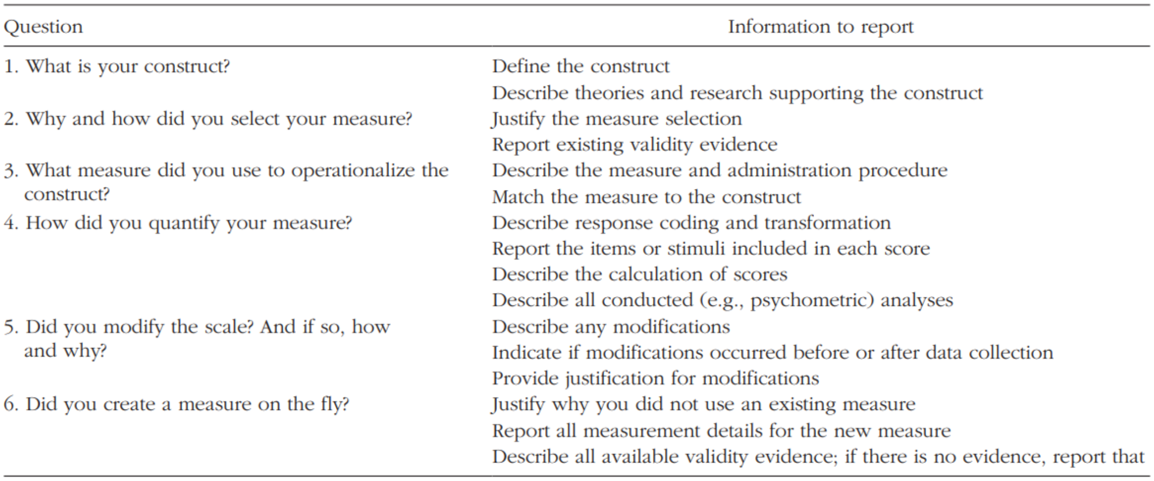
Defining and measuring constructs is an underlying aspect of science (Flake & Fried, 2020). However, this fundamental practice has been overlooked especially in the field of social science (Schimmack, 2021). As most constructs in this field cannot be observed directly (Flake & Fried, 2020), the chance of discrepancy among scholars about how to define and measure these constructs is high. For example, some scholars define internet addiction as an impulse control disorder, while others classify it as a behavioural addiction, and even some use a combination of these two definitions (Abendroth et al., 2020). In another example, anhedonia is sometimes assessed as a neurological dysfunction associated with schizophrenia and is sometimes considered a premorbid personality trait that predisposed patients to the development of schizophrenia spectrum disorders (Winer et al., 2019). But how can we measure a construct irrespective of its different conceptualizations? Even a noted construct such as depression still suffers from validity concerns about its definition and measurement (Flake & Fried, 2020; Fried & Flake, 2018). This is a problem that neither comprehensive design nor statistical tests can solve; this is why it is important to pay special attention to the validity.

Overall, there are of four types of validity. *Internal validity* refers to the degree to which we can draw causality from a study; *External validity* addresses the generalizability of the results drawn from the study; *Statistical validity* captures whether the data supports the conclusion or not; and *construct validity* concerns the operationalization of the constructs (Moring, 2017). Each type of validity can be threatened by measurement. For instance, if necessary information regarding the operationalization is missing, construct validity is threatened, or the absence of information regarding whether the measure is sample- or population-specific can endanger external validity (Flake & Fried, 2020). The term validation crisis refers to the widespread use of *questionable measurement practices* (QMPs). According to Flake & Fried (2020), QMP is defined as any decisions researchers make that increase uncertainty about the validity of measures, and in general, the validity of the final claim.

Unfortunately, many psychologists are satisfied with reliability as the sufficient criterion of validity or fail to report validity as they know it is embarrassingly low in the field (Schimmack, 2021). Therefore, it is essential to consider the validation crisis seriously, as without valid measures, significant and replicable results will be uninformative or even wrong (Schimmack, 2021). One way to make sure measurement leads to more robust insight is to compare the results of different scales of the same construct (Fried & Flake, 2018). The idea of this approach is as different measures can lead to different conclusions by comparing the results drawn from each scale, we can mitigate the results. However, this approach leaves room for questionable research practices such as p-hacking (Fried & Flake, 2018).

As another approach, Flake & Fried (2020) suggested researchers transparently justify the use of a certain measure. To do so, they proposed a set of questions (Figure 1.2) about the validity of the study and encouraged researchers to use these questions as a guide when they are designing a study (Flake & Fried, 2020). These questions can also enable reviewers, editors, and consumers of research to detect QMPs and have a more informed evaluation of the research (Flake & Fried, 2020).

On the other hand, Schimmack (2021) suggested the only way to study validity (specifically construct validity) is by quantifying it and using casual modeling with SEM. He emphasized that validity cannot be dichotomized, rather it is a dynamic process through which the variance of validity can change as more information becomes available (Schimmack, 2021). However, this approach is not without limitations; some areas of psychology lack the necessary theoretical background and therefore, do not provide enough information for structural equation modeling (Schimmack, 2021); It also does not provide justification for the cut-off point to reject the validity of the measurement. In general, it is important to consider the overlooked impact of the credibility crisis.



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: Proposed questiones to enhance transparency of reporting measurment practices Adapted from Measurement schmeasurement: Questionable measurement practices and how to avoid them by J. Flake and E. Fried, 2020, Advances in Methods and Practices in Psychological Science, 3 (4), P. 459

Psychologists have procrastinated to address this problem for so long, and this field needs more research to find more practical ways to mitigate the effects of this practice.

1.2 Possible remedies for credibility crisis

In previous sections, we discussed the negative impacts of the credibility crisis and the practices contributing to it. In this section, we will focus on more practical ways to prevent the impact of these practices on scientific literature and public trust.

One of the celebrated remedies for the credibility crisis is the “Open science” movement. Open science is an umbrella term referring to a variety of practices and principles to ensure transparency, credibility, reproducibility, and accessibility (Kathawalla et al., 2021). Transparency can cover different aspects; free online access to articles through open-access publishing, providing access to data, codes, and methods, providing incentives to disclose more of the results, and making the peer review process more transparent (Elliott, 2022). Transparency can also take different degrees. Conceptually, the degree to which aspects of a project can be disclosed depends on the purpose of transparency and the needs of the audience (Elliott, 2022). For example, if the primary purpose of transparency is to enable others to reproduce data, then the audience for it would be other scholars, and they need all the information including codes and data to reproduce the results. Although transparency increases the probability of reliability and reproducibility and helps to accelerate scientific innovation, it can also give information to those who want to harass scientists (Elliott, 2022). These harassments varies from complaints to researchers' universities to threats of violence (Lewandowsky & Bishop, 2016). We should notice that all legitimate tools can be abused; however, by extending protective actions and raising awareness about these harassments, we can hope to alleviate their damage to the lowest degree (see Lewandowsky & Bishop, 2016). One of the tools that can satisfy some aspects of transparency is *preregistration*. Preregistration refers to a dated document containing research questions, the hypotheses, the method, and the analysis plan that is published before data collection (Kathawalla et al., 2021). The idea is that preregistration stops intentional or unintentional decisions that can affect the outcome (M. Bakker et al., 2020). Preregistration also enables researchers to receive feedback about their design and as a result, prevents impractical outcomes drawn specially by expensive studies that are unlikely to replicate (Elliott, 2022). Additionally, it leads to improvement in study design in a way that by looking at the checklist, researchers will be reminded of some details that otherwise have been overlooked (Lakens, 2019). Preregistration also helps others to evaluate the extent to which a prediction can be falsified (Lakens, 2019). However, it is important to mention that preregistration itself does not increase the value of the study compared to a non-preregistered one (Lakens, 2019), and also one template of preregistration does not work for all kinds of research (Miguel et al., 2014). Preregistration also has received some criticism. Some scholars argue about the role of preregistration in suffocating exploratory research and creativity (Miguel et al., 2014). Nevertheless, the aim of preregistration is to make sure the exploratory analysis is not going to be portrayed as formal hypothesis-testing, and by no means this practice would damage exploratory findings (Miguel et al., 2014). From another perspective, Van Rooij (2019) mentioned that focusing on preregistration prevents addressing the deeper problem of theory development. I do not see how registration can prevent theory development, as this "theory crisis" has roots in the validity crisis and weak evidence (Eronen & Bringmann, 2021). Theory crisis in psychology has many contributing factors that are not in the scope of the current study; however, redirecting attention from one crisis toward another does not seem to be reasonable. Furthermore, Szollosi et al. (2020) raise concerns that poor theories can also be preregistered and how *post hoc* inferences are overlooked just because they were not thought of before preregistration. As we mentioned earlier, preregistration itself does not imply that a study is well-done or has a good theory. Moreover, registration, in my view, does not prevent *post hoc* inference, it simply differentiates the tests that have been decided before data collection and looking at the data.

The Open Science Framework (OSF) is one of the tools to enhance transparency during different steps of the research lifecycle (Foster & Deardorff, 2017). The main function of OSF is to make and develop projects, which can vary from a particular paper to the work of the entire lab (Foster & Deardorff, 2017). This framework facilitates collaboration and also has a feature to make the entire projects, or some aspects of them, publicly available (Foster & Deardorff, 2017). OSF also offers different preregistration formats, that once get registered, cannot be edited or deleted (Foster & Deardorff, 2017). Even if you withdraw a project, a record of the registered project still remains (Foster & Deardorff, 2017). We are not implying that OSF is the best tool for transparency; however, it helps researchers to perform many open science practices and does not involve learning many interfaces (Kathawalla et al., 2021).

As we mentioned in the previous section, many questionable practices are performed to increase the chance of publication. The *registered report* (RR) is an approach to specifically address this issue for hypothesis-driven studies. The idea of RR is that the study proposal gets peer reviewed and accepted before the actual study is done (Chambers & Tzavella, 2022). By doing so, the blind focus on the results of the study would shift toward question, theory, and methods (Chambers & Tzavella, 2022). In this approach, peer review is divided into two stages (Figure 1.3). In the first stage, authors write in detail the introduction, methods, and analysis plans and submit it for peer review; the final outcome for favourably assessed proposals is "in principle acceptance" (IPA), which means the journal guarantees to publish the final paper if the authors follow their peer-reviewed protocol (Chambers & Tzavella, 2022). After the research is done, the authors add results and discussion sections to the previously approved protocol and submit it again, so reviewers can check whether the protocol was followed and whether the evidence justified the conclusion (Chambers & Tzavella, 2022). Although currently, more than 300 journals offer RR, we should remember that registered reports were introduced in 2012; therefore, we cannot say with confidence that this practice reduces bias and improve reliability, and of course, there are many aspects of this approach yet to be developed (Chambers & Tzavella, 2022). Nevertheless, there is evidence for signs of bias control, study quality, computational reproducibility, and citation influence (see Chambers & Tzavella, 2022). Changing the mindset about the significant result is a must, and we still have a long way to reach there. But surely RR is a practical tool to train researchers and journals to change their view about worthy results.

Alongside being more transparent, researchers should be encouraged to engage more in replication. One study is barely enough to jump to a conclusion. It is by replication that we can proceed toward increasing knowledge and generating a scientific law (Chambers, 2017) instead of sticking to this "vast graveyard of undead theories" (Derksen & Morawski, 2022). One way to incentivize researchers to perform replication studies is that journals guarantee publication of the well-done replications (Chambers, 2017). Another proposal is that journals would be obligated to publish direct replications of

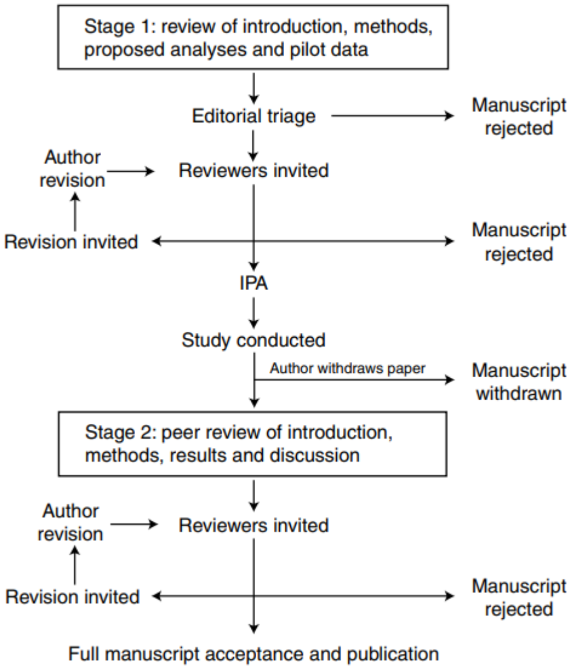


Figure 1.3: The process of RR Adapted from The past, present and future of registered reports by C. Chambers and L. Tzavella, 2022, Nature Human Behaviour, 6(1), P. 30

any original findings (Chambers, 2017). The role of journals in normalizing replication is undeniable (Chambers, 2017). Unfortunately, still journals are not prone enough to publish replication studies (Martin & Clarke, 2017). Moreover, certain restrictions of space even in online journals are preventing authors to put sufficient information about the methodology and statistical analysis of the study, which is essential for replication (Chambers, 2017). Therefore, certain changes in policies should be implemented.

In the previous section, we also refer to how different decisions can contribute to the credibility crisis. Generally, these choices are referred to as the *researcher's degree of freedom* which we will address in more detail in section 2.1. There are several approaches to mitigate the impact of these choices on the result of the study. Among those are the *multi-analyst* approach and the *multiverse analysis*. The former is based on the idea that different scholars take different decisions (Aczel et al., 2021). To have an informed idea of how the results are influenced by these decisions, one should employ several analysts from different labs to independently run analyses on the same data set (Aczel et al., 2021). The idea is that by evaluating several independent analysis options, we can systematically assess whether conclusions are dependent on the researcher's degree of freedom (Aczel et al., 2021). The latter follows this mindset that we should increase transparency to the fullest and consider *all* plausible choices when we are assessing their impact on the final results (Steegen et al., 2016). In the next chapter, we discuss this viewpoint in more detail.

1.3 Aims of the current study

In this chapter, we discussed the importance of transparency and its role in the credibility crisis. There is usually great flexibility in analytical choices and data preprocessing which is partly due to imprecise theory and hypotheses (Hoffmann et al., 2021). These multiplicities would endanger the credibility of science by increasing the chance of non-replicability (Hoffmann et al., 2021). One of the aforementioned approaches to tackle this problem is to consider all arbitrary yet plausible decisions at each level of the research process, which is termed the multiverse analysis (Steegen et al., 2016). Multiverse analysis benefits the credibility of science by transparently depicting how results are dependent on decisions. One ground to plant the idea of the multiverse is in meta-analysis. By combining single studies, meta-analysis enables us to estimate a parameter more accurately and with sufficient power (Maxwell et al., 2008). However, it requires certain decisions which would potentially affect the final outcome of the meta-analysis, such as choice of the meta-analytic model, inclusion and exclusion of potential outlier studies, and imputation of statistical information that often are not reported in the original studies but is needed to model the observed data (Voracek et al., 2019). Therefore, combining these two ideas would benefit the audience to have an informed understanding of the heterogeneity of the results based on the combinations of options, with the ultimate goal to deeply evaluate the robustness (or fragility) of research findings. One challenge in combining these ideas is that there is no specific framework to present the outcome of multiverse meta-analysis. In this study, we aim to propose an exploratory framework to present and depict the result of multiverse meta-analysis in a way to ease the interpretation of the research findings. In addition, we would highlight the influence of each decision on this heterogeneity.

Chapter 2

Multiverse Analysis

2.1 What is multiverse analysis?

As we mentioned in the previous chapter, choices made by the researcher can contribute to the credibility crisis. Empirical research requires scholars to make certain decisions. However, these choices, especially choices concerning methodology, are often made randomly and with no justification (Wicherts et al., 2016). Simmons et al. (2011) termed these decisions as the *Researcher Degree of Freedom* (RDF). Generally, the researcher degree of freedom is believed to have a negative implication (Gelman & Loken, 2013). The reason is that it is believed that researchers take these decisions deliberately to increase the probability of having significant results (which usually are false positives) or to inflate effect sizes (Wicherts et al., 2016). This definition of researcher degree of freedom implies that researchers continuously get involved with questionable research practices to get the most desirable outcome from a data set (Gelman & Loken, 2013). This may be the case for some researchers, as we mentioned in the previous chapter, however, it is not the whole truth. Those who do not engage in multiple different analyses are still subjected to problems of RDFs (Gelman & Loken, 2013). In this study, by using the term researcher degrees of freedom we mean consciously or unconsciously choosing one option among multiplicities at any level of study that can potentially affect the results of the study. Given this definition, all aforementioned questionable practices as well as common data processing and analysis decisions can fall in the category of researcher degrees of freedom.

As we mentioned, the researcher degree of freedom has the potential to affect the outcome. Therefore, with each decision, researchers make a turn in the maze of possibilities and face another crossroad in that maze. Gelman & Loken (2013) term this process as the *garden of forking paths*. Figure 2.1 shows a simple case where at each level of study, researchers have two options. Sometimes, no matter which choice they make, all paths lead to the same conclusion. This may be the case where there are large real differences, small measurement errors, and low variation (Gelman & Loken, 2013). However, these prerequisites are hard to find in a field like psychology. Expectedly, we can see that sometimes different paths reach different conclusions (Fig. 2.1). In other words, by choosing one path (the thick black line), we only reach one conclusion that might not even be the most probable one.

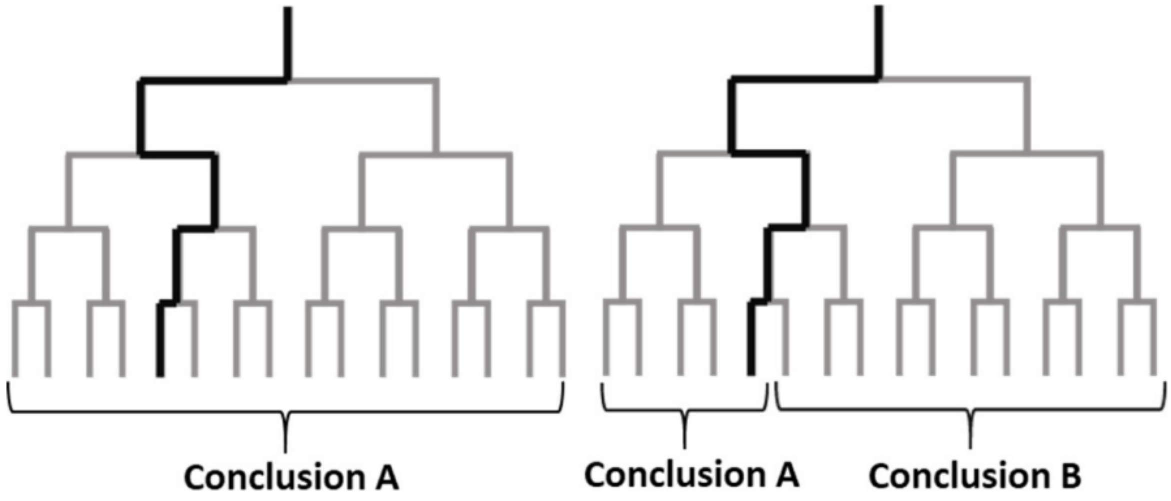


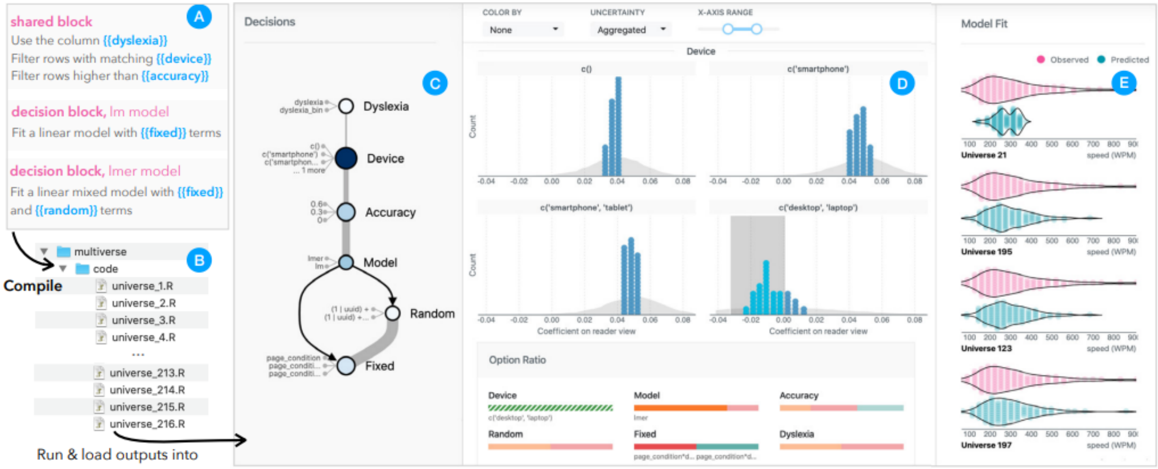
Figure 2.1: The garden of forking paths by B. Aczel et al., 2021, eLife, 10, P. 3

It is evident how this practice can attribute to the credibility crisis. There is already some evidence in the literature directing toward the problem of different and sometimes opposite conclusions which is derived from taking specific decisions (Aczel et al., 2021; Gelman & Loken, 2013; Modecki et al., 2020). Arbitrary choices in statistical analysis and model choice such as repeating measures ANOVA vs. using linear mixed models, assuming normality, and choosing between parametric and non-parametric approaches can affect the result of the study (Steegen et al., 2016). In one of the most cited examples, Silberzahn et al. (2018) gave the same data set to 21 research teams to independently check whether football referees are more likely to give red card to dark-skin-toned players, which resulted in 29 different analyses with 21 unique combinations of covariates. Not only the effect size estimates were highly dispersed, but 31% of teams also obtained a null effect, which in two cases was even numerically negative. This case perfectly shows the effect of choosing different analyses on the obtained results, where neither the prior belief of the researcher nor the researcher's level of expertise could explain the variation (Silberzahn et al., 2018). As we mentioned before, RDFs do not limit to statistical analysis, but they also involve data processing. Preparing data for analysis is not a passive act, rather converting raw data to a form suitable for analysis can be considered as a data construction (Steegen et al., 2016). In this process, scholars engage in many RDFs, such as deciding on the categorization and dichotomization of variables, combining variables, and exclusion of data (Steegen et al., 2016). We already know that by dichotomizing a quantitative measure, we lose information and misestimate the effect sizes (MacCallum et al., 2002). However, different kinds of dichotomization of categorical variables also affect the results. In their paper, Steegen et al. (2016) showed a case where using a different dichotomizing method for fertility and relationship status affected the interaction effect of relationship status on the relationship between fertility and religiosity. This is evident that choosing only one path may be deceiving (Heyman & Vanpaemel, 2022), but the question is how can we draw an informed conclusion if the results of our study are so dependent on the decisions we made?

Steegen et al. (2016) proposed the **multiverse analysis** to answer this problem. They argued that for each study we have a multiverse of data sets, statistical analyses, and models, as well as a multiverse of outcomes. By choosing one data set or analysis and ignoring the other possible and yet reasonable options, our results are in danger of being fragile with no way to evaluate the robustness of our findings (Steegen et al., 2016). As a result, in the absence of a precise and complete theory, which clarifies why one option is better than the others, they proposed to analyze all reasonable combinations of choices we have in the garden of forking paths (Steegen et al., 2016). In the complete format of the multiverse analysis, one crosses the multiverse of data processing with the multiverse of analytical decisions to consider all possible combinations (Steegen et al., 2016).

> The multiverse analysis is a method to consider all plausible options in different levels of the study (e.g., data processing or statistical analysis) to evaluate the robustness of the results.

It is evident that multiverse analysis requires great effort and time to handle a multiverse of sets. To facilitate this hustle, some R packages have been proposed (e.g. Masur & Scharkow, 2020; Sarma et al., 2021). However, no matter how we run the multiverse analysis, we will end up having a lot of results, which need to be managed and depicted. Steegen et al. (2016) used several histograms and matrixes to show multiple *p*-values drawn from multiverse analysis. In another study, Modecki et al. (2020) used panels of the scatter plots to depict *p*-values and effect sizes for predictions of parenting by smartphone usage variables. They also used bar plots to show moderator effects, and they ran a meta-analytic sensitivity check to realize which model attributed the most variation in effect size (Modecki et al., 2020). Lately, a programming tool called Boba (Figure 2.2) has been proposed to facilitate conducting and visualizing the multiverse to better assess the result of all paths (Liu et al., 2021; Liu, 2022). One can use the method they consider the best as long as their choice transparently shows all results of the analysis.



shared block
Use the column {{dyslexia}}
Filter rows with matching {{device}}
Filter rows higher than {{accuracy}}

decision block, lm model
Fit a linear model with {{fixed}} terms

decision block, lmer model
Fit a linear mixed model with {{fixed}} and {{random}} terms

COLOR BY: None
UNCERTAINTY: Aggregated

Figure 2.2: Conducting and visualizing multiverse analysis with Boba by Y. Liu et al., 2021, IEEE Transactions on Visualization and Computer Graphics, 27(2), P. 1753

Many researchers proposed ideas similar to the multiverse analysis. Young & Holsteen (2017) by proposing a multi-model approach focused on the role of model assumption on the results. Researchers must take necessary choices on model assumptions to use a certain model for concluding the result. However, only under two conditions, one point estimates would be enough for covering the multiverse of estimates (Young & Holsteen, 2017). The first condition is that the researcher knows the true model and overlooks other models because of their inaccuracy and misleading potential. The second one is when the researcher is confident that all plausible models yield similar conclusions (Young & Holsteen, 2017). It is evident that none of the assumptions can be met. Therefore, to understand the robustness of the result, they proposed to model the distribution of possible estimates across all possible combination of the model assumption and check how each model component influence the coefficient of interest (see Young & Holsteen, 2017). An important drawback of this approach is that, unlike traditional model averaging, this approach does not weigh models differently according to their fit, and this can result in an overestimation of the uncertainty drawn by model choice (Slez, 2019).

As another approach, Simonsohn et al. (2020) emphasized that when one chooses only one analytic option in the garden of forking paths, the standard error cannot reflect the error caused by researcher degrees of freedom. Therefore, by proposing the specification curve analysis, they aimed to minimize the effect of these decisions that are not driven by theory or prior beliefs (Simonsohn et al., 2020). The idea of specification curve analysis is quite simple; if there are several reasonable analyses to test the research question, and all are statistically valid and are not redundant with other analyses, we should run them all, summarize them in a curve plot (Figure 2.3), and evaluate the result across all of them (Simonsohn et al., 2020). This approach is the first attempt to draw inferential conclusions from previously exploratory multiverse results (Girardi et al., 2022) as it enables researchers to compare the result of the multiverse with a null distribution made by bootstrapping or permutations (Srivastava, 2018). However, it is not without limitations. Some analyses in certain cases may be superior to other theoretically justified and valid options. However, specification curve analysis cannot weigh them differently (Simonsohn et al., 2020). Additionally, it does not allow testing all possible specifications (Rauvola & Rudolph, 2023; Simonsohn et al., 2020), as well as it can only be run on simple cases related to the linear model (Girardi et al., 2022). The last proposal we cover here is the multi-analyst approach. This approach was proposed more than a century ago, however, its use has not been widespread among many scientists (Aczel et al., 2021). The idea of this approach is that instead of exhaustively evaluating all sensible analyses, we can check the robustness of the result by checking the analyses of several analysts (Aczel et al., 2021). As in the aforementioned example, independent analysts choose different specifications and analyses, and then, it would be evident how much the conclusions are dependent on the analytical paths they have chosen (Aczel et al., 2021; Silberzahn et al., 2018). One limitation of this approach is that when the research question is not precise enough, which unfortunately is the case in many social science studies, using this approach leads to an overestimation of

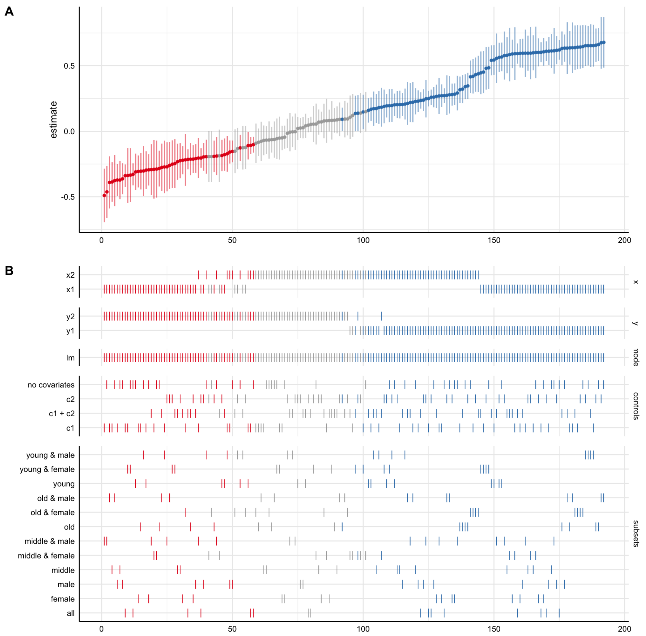


Figure 2.3: Specification curve plot by N. Ballou and A. Van Rooij, 2021, R. Soc. Open Sci., 8, 12

the variability of the results (Auspurg & Brüderl, 2021). It is also not clear how the number of analysts should be justified; and how the final report should be written in case of disagreement among analysts (Auspurg & Brüderl, 2021).

Multiverse analysis has several advantages. This analysis improves transparency to a considerable extent (Steegen et al., 2016). In the previous chapter, we discussed the benefits of transparency and its role to improve the credibility crisis. However, multiverse analysis goes one step further than simple preregistration (Steegen et al., 2016). Multiverse analysis not only reveals and stops questionable practices but also shows how the results of a study are changing as a function of decisions (Pipal et al., 2022). Some might raise the issue that multiverse analysis can itself provide a tool for those who want to engage in questionable research practices (Akker et al., 2021; Masur & Scharkow, 2020). Although this is a genuine concern, we should remind that most innovations have the potential to be used for harmful goals. Additionally, if the study is preregistered, then one cannot use multiverse analysis as a way for p-hacking or cherry-picking. Although preregistering the multiverse is not as straightforward as a normal analysis, it is not impossible. In this link (https://osf.io/mcs8r), you can find an example of preregistration done for multiverse analysis (Wessel et al., 2020). As it is shown, the researcher should include all information about the number of tests, the number of data sets, how to define the outliers, etc. Another important benefit of multiverse analysis is that it can be used in several contexts. Multiverse analysis can be used with both frequentist and Bayesian frameworks (Dragicevic et al., 2019; Haaf et al., 2020; Liu, 2022). It can also be applied in the context of meta-analysis (Voracek et al., 2019). The idea of multiverse analysis can also be combined with explorable explanations which culminate in reporting multiverse in a way that reader can dynamically move through the alternative analysis options (Dragicevic et al., 2019). Its characteristics also make multiverse analysis a perfect candidate for students' research projects (Heyman & Vanpaemel, 2022). Not only it addresses the serious problem of lack of evidence for the robustness/fragility of results in the literature, but it also helps students to practice multiple statistical analyses, and it saves them time in gathering data (Heyman & Vanpaemel, 2022).

Despite all benefits, we should not forget about the disadvantages of this approach. Multiverse analysis is highly context-specific (Steegen et al., 2016). Alternative options vary considerably depending on operationalization (Hanel & Zarzeczna, 2022), measurement (Harder, 2020), research question (Auspurg & Brüderl, 2021; Steegen et al., 2016), and researchers who are performing the study (Steegen et al., 2016). This subjectivity can affect the replicability of multiverse analysis negatively. Moreover, although multiverse analysis has the advantage of considering all plausible ways, it does not weigh these options differently based on theory. Steegen et al. (2016) response to this problem was to use the knowledge from theory post-hoc to interpret the results. This response did not convince me as properly formalized theory is like having a wider path in the garden of forking paths compared to the narrow alternatives. However, it does not apply one should be substituted over the other (Krypotos et al., 2022). In addition, while in theory, as we mentioned before, the most comprehensive way to do multiverse analysis is to consider the multiverse of data processing and multiverse of analytical decisions at the same time, in practice only one is possible (Hoogeveen et al., 2022). This is because of restrictions in both interpretability and practicality (Hoogeveen et al., 2022). Furthermore, few available references are guiding how to perform multiverse analysis (Rijnhart et al., 2021). This can explain why the number of studies using multiverse analysis is still low (Rijnhart et al., 2021). Adding to that, those who are using multiverse analysis mostly stuck to descriptive reports rather than inferential conclusions (Girardi et al., 2022). This may be due to the scarcity of approaches to make inferential decisions from multiverse analysis. We mentioned specification curve analysis as the first effort to draw inferential conclusions before, and we stated its drawbacks. Recently, Girardi et al. (2022) have proposed the *Post-selection Inference approach to Multiverse Analysis* (PIMA), which is much more flexible and can be used for a wide variety of models. This proposal may encourage researchers to use multiverse analysis more, yet, as it is particularly a new approach, evidence to explore its practicality should yet be collected.

2.2 Adaptation of the multiverse

The idea of multiverse analysis has become popular among researchers in the past few years (Hanel & Zarzeczna, 2022). However, some scholars noticed that multiverse analysis has the potential to get implemented in different contexts. In this section, we will discuss three of these new proposals: **multiverse operationalizations**, **multiverse of methods**, and **multiverse meta-analysis**.

Most examples we covered so far focused on different decisions researchers take after data collection. In other words, multiverse analysis has been mostly done on only one set of raw data (Harder, 2020). However, decisions during operationalization and measurement that happens before data collection can also impact the result of the study (Hanel & Zarzeczna, 2022; Harder, 2020). Different scales used to operationalize the variables can affect the criterion validity and the result of the study (Hanel & Zarzeczna, 2022). In one example, the result of the longer scale for the need for cognition (NfC) was twice as much correlated to reliance on policy information than the shorter version (B. N. Bakker & Lelkes, 2018). Additionally, the MARP team (2022) showed different operationalizations can affect the outcome of the study. Although they did not find the impact as large as the effect of analytical decisions (Hoogeveen et al., 2022), it is important to notice decisions at the measurement level potentially can change the outcome of the study. The idea of *multiverse operationalizations* is to consider every possible way of operationalizing the construct and compute the results across all of them (Hanel & Zarzeczna, 2022). In their study, Hanel & Zarzeczna (2022) documented how using fewer items in defining a construct culminates in a wider spread of findings. Although these results are enlightening, Hanel and Zarzeczna did not take into account the effect of using different scales. Although we expect using different measurements can contribute to the dispersion of results, there is considerable difficulty in studying this effect. As we mentioned in the previous chapter, psychology is indeed suffering from the validity crisis, and assessing the impact of equally valid measures on the result of a study may not yet be achievable.

As another expansion proposal, Harder (2020) tried to direct our attention toward the impact of using only one raw data set on the multiverse of results. She stated that the lack of clarity can be expanded to how to implement measurement as well as the participants of the study, and as a result, considering only one set of raw data will not give us the whole image of the multiverse. In doing so, she proposed the *multiverse of methods* in which she considered the multiverse as all studies on the same phenomenon that are varying on data-collection methods (Harder, 2020). The *shooting bias* example can perfectly depict the idea of this proposal. In these sets of studies, researchers want to depict the racial bias in the decisions of police officers on whether to shoot a suspect (Correll et al., 2014). One task to measure such bias is called the first-person-shooter task (FPST), which presents several pictures of either dark-skin-tone or light-skin-tone males carrying a gun or a harmless object, and participants should choose between pressing a key labeled “shoot” or the other key labeled “don’t shoot” (Correll et al., 2014). However, different analytical and methodological decisions potentially have an effect on altering the results drawn from such studies. These studies have two outcome variables, errors and reaction time, yet, there is no general agreement in choosing which variable to measure this bias (Harder, 2020). The second variation is how to analyze the error data, as the number of observations and their reliability can affect the overall error rate (Harder, 2020). Another variation refers to the pool of stimuli in the study. White targets may be less muscular or have other traits to be perceived as less threatening, which some studies do not take into account (Harder, 2020). Another source of difference is that these studies present a different number of targets to participants and allow participants to have a varied time limit to press the button at each trial (Harder, 2020). Although traditional multiverse analysis only considers analytical decisions in this example, Harder (2020) showed that methodological decisions play the role of moderators in the study. This idea opens the door for further exploration and adaptation of the multiverse idea, however, it also faces some limitations. The multiverse of methods is bounded by published literature (Harder, 2020). This can cause two problems; first, one does not have a clear image of all possible combinations of methodological decisions because of the few numbers of studies and/or publication bias (See Section 3.5). Second, in some cases, the results from the at-hand multiverse of studies may be missing as a result of convergence issues (Harder, 2020). Additionally, the multiverse of methods is subjected to a hidden variation in data sets that cannot be controlled, such as the effect of the lab of origin on the data (Harder, 2020). In any case, the idea of the multiverse of methods can broaden our perception of how the results are under the influence of the decisions taken by the researcher.

Another ground to expand the idea of the multiverse is in meta-analysis. Meta-analysis is a tool to systematically review individual studies while giving weight to each study based on prespecified mathematical criteria (Borenstein et al., 2009). Despite many benefits of meta-analysis, it involves several degrees of freedom (Voracek et al., 2019), such as inclusion criteria for studies, meta-analytic modeling, and imputation of relevant statistical quantities that may be missing in some of the considered studies. If we do not take the ambiguity of these necessary decisions into account, the result of the meta-analysis might mislead us. Therefore it is the perfect ground to plant the multiverse idea. Voracek et al. (2019) adopt specification curve and multiverse approaches into meta-analysis and introduced the term *multiverse meta-analysis*. The role of multiverse meta-analysis is to contain all theoretically and conceptually justified meta-analyses that can be conducted on a research question (Plessen et al., 2022; Voracek et al., 2019). Voracek et al. (2019) then proposed to test the results using the inferential statistical test from specification curve analysis and display them graphically using a mixture of proposals from multiverse and specification curve analyses. This proposal is not the only one for doing the multiverse meta-analysis. As drawing inferences from the multiverse is still a matter of controversy, one may want to increase transparency by descriptively checking how much meta-analytic results vary if they choose different paths (e.g. Donnelly et al., 2019). In the next chapter, we will discuss meta-analysis in more detail, and explain different compulsory decisions that should be taken while doing a meta-analysis.

Chapter 3

Meta-Analysis

> "... [in the meta-analysis context,] a researcher can never be sure what the true underlying model (fixed-effect or random-effects) is for the sample effect sizes collected from many studies; as a result, one can never be certain which (fixed-effect or random-effects) is the model consistent with the nature of the sample data."
>
> — Cai & Fan, 2020, p. 13

3.1 Introduction to Meta-Analysis

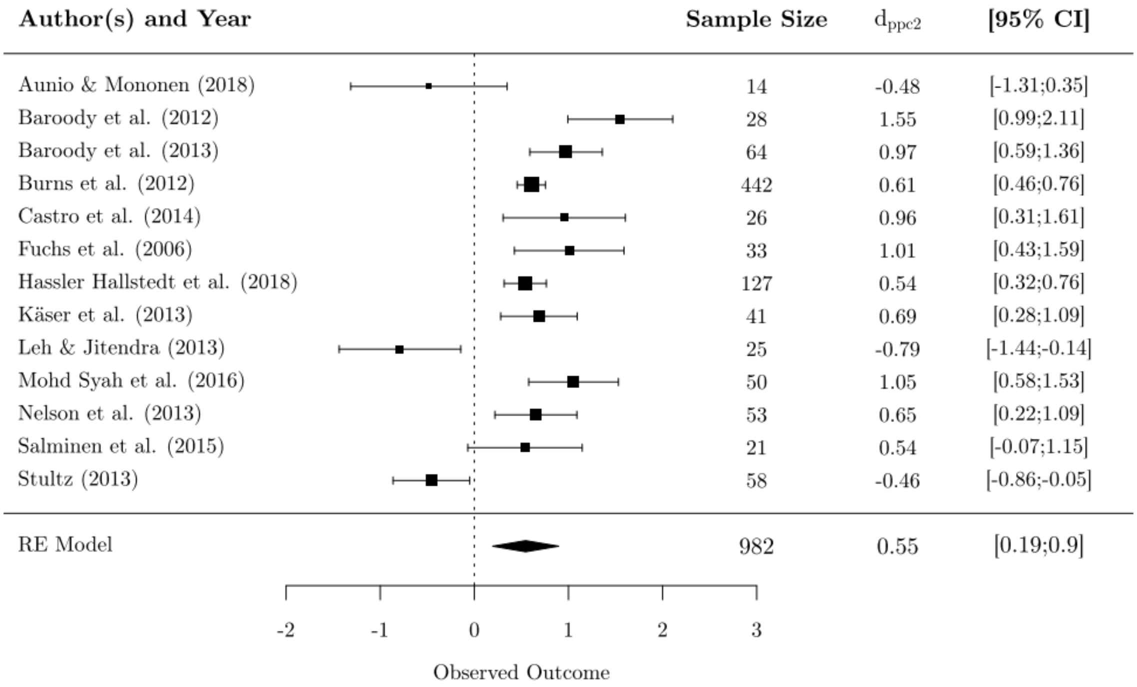
The number of published studies has increased exponentially in the past decades (Harrer et al., 2021). Although this trend may be exciting for many, there is a concern about the stable and perpetuated fallacies that are re-accruing in generations of studies (Harrer et al., 2021). This problem could not be tackled by narrative reviews which used to be the only option for summarizing and synthesizing the results of the pile of studies, as it neither offers any uniform criteria for assessing the studies nor was its results useful once more evidence was at hand (Borenstein et al., 2009). This need was met by introducing *systematic reviews*. This method was developed to not only synthesize the results of a great number of studies but also to generate robust results that are as unbiased as possible (Mallett et al., 2012). Systematic reviews can be performed on both quantitative and qualitative studies, as long as the reviewer follows the predetermined and transparent rules of the systematic reviews (Harrer et al., 2021). However, when the aim of the systematic review is to integrate the quantitative outcomes of studies into one numerical estimate, it is called *meta-analysis* (Harrer et al., 2021).

Meta-analysis has gained more popularity over the past few years in social, behavioral, and health sciences (Cai & Fan, 2020). Scientific journals are often more prone to publish meta-analyses and these studies are often more cited (Harrer et al., 2021; Polanin et al., 2020). Therefore, it is essential for meta-analyses to be as high-quality as possible. Before performing any meta-analysis, it is important to ask when is it logical to perform a meta-analysis and what studies should be included in the analysis (Borenstein et al., 2009). These are valid questions as many single studies on similar research question are different from one another in various ways. As we highlighted before, single studies barely have sufficient power to estimate the statistical parameters of the population (Maxwell et al., 2008); However, combining studies without justification would probably culminate in meaningless results (Borenstein et al., 2009). To understand whether studies are similar enough to be compared quantitatively, we should consider the question we want to answer (Borenstein et al., 2009). If our question is to check the efficacy of a certain drug to reduce symptoms of ADHD in children aged 8-15 years old, we cannot add studies performed on other age groups. In addition, we need to choose studies that have designs to assess this efficacy more precisely. Sometimes theoretically we need to use experimental studies, but those at hand are poorly done, and as a result, they will not be the best candidates to run a meta-analysis on.

Meta-analysis has faced considerable criticism over the years, and some are still difficult to deal with. Many meta-analyses, especially in the field of psychology, cannot be reproduced (Lakens et al., 2017). Although reviewers are improving in reporting transparently, many have neglected to report some methodological elements, such as necessary information on estimating the effect size and enough information about studies involved in the meta-analysis (Polanin et al., 2020). In order to improve the reproducibility of the meta-analysis, one can take several practical steps, like pre-registering the meta-analysis, disclosing all meta-analytic data, and adhering to reporting results according to standards such as PRISMA (see Lakens et al., 2016). Another criticism is referred to as *garbage in, garbage out*. Based on this criticism, when we include studies that are biased or low-quality, the resulting meta-analysis is equally flawed (Harrer et al., 2021). As meta-analytic studies have more impact on the literature, doing such biased meta-analyses would do more harm. Another common criticism is that meta-analysts combine different kinds of studies in a single analysis, which is referred to as the *mixing apples and oranges* (Borenstein et al., 2009). Although meta-analysis can calculate numerical estimates regardless of the studies involved in the review, the estimate will be meaningless if studies do not share the necessary characteristics to answer a certain research question (Harrer et al., 2021). It is important to note that when meta-analysis is to be performed on literature, studies are inevitably different from one another, and it is the reviewer's task to decide how similar they should be (Borenstein et al., 2009). Therefore, this problem highly depends on the question the researcher tries to answer (Harrer et al., 2021). In any case, meta-analysis can investigate the impact of these differences on the outcome if it is required (Borenstein et al., 2009). Lastly, performing a meta-analysis of the literature is inevitably biased. We discussed the intolerance of journals against null results (more generally, unsatisfactory results) in chapter 1, and we will look at this phenomenon in more detail in the subsequent section on publication bias (see section 3.5). This selective bias results in what is called the *file-drawer* problem, where researchers would not/could not publish their "disappointing" results. As published results always have more chance to be involved in the review (Borenstein et al., 2009), meta-analysis reflects this biased view of the literature and misestimates the effect size in the end (Borenstein et al., 2009; Ter Schure & Grünwald, 2019).

On the other hand, the advantages of a well-performed meta-analysis will surpass its disadvantages. As meta-analysis combines several studies, hence a greater sample size, it has more power and accuracy to study an effect (Crocetti, 2016; Maxwell et al., 2008). Additionally, not only it can address theoretical research questions but also can investigate methodological issues such as the reliability of an instrument (Crocetti, 2016). Meta-analysis also differentiates the studies in the pool by giving weight to them (Borenstein et al., 2009). These weights can represent specific goals, such as minimizing the variance or reflecting the range of effect sizes (Borenstein et al., 2009). Another advantage of this method over other reviews is that it takes into account the dispersion in the results and explains the emerged patterns between studies (Borenstein et al., 2009). Moreover, some questionable research practices (e.g., p-hacking) are less prevalent in the meta-analysis (Voracek et al., 2019). Lastly, meta-analysis can adopt the Bayesian approach (Sutton & Abrams, 2001). Using a Bayesian framework in meta-analysis allows accounting for all parameter uncertainty and extension of models to adjust more complex scenarios (Sutton & Abrams, 2001).

Before moving further in this chapter, it is necessary to also have an overview of how to visualize the result of a meta-analysis. The *forest plot* is a common and quick way to understand the result of a meta-analysis and learn which studies are included in it. As you can see in figure 3.1, a forest plot consists of three main parts. On the left, all studies included in the meta-analysis are specified. The summary of their results is depicted by a square in the central section within its bounds of confidence. The size of the square is an indicator of how much weight has been given to that study in the meta-analysis. On the right side of the figure, the numerical values of the results are presented in their specified column. The last row, which usually is divided by a line from the rest of the rows, dedicates to the synthesized result of the meta-analysis. It is presented graphically in form of a diamond. The center of the diamond represents the pooled effect size, and the two tails represent the confidence interval around that effect size.



Author(s) and Year	Sample Size	d_{ppc2}	[95% CI]
Aunio & Mononen (2018)	14	-0.48	[-1.31;0.35]
Baroody et al. (2012)	28	1.55	[0.99;2.11]
Baroody et al. (2013)	64	0.97	[0.59;1.36]
Burns et al. (2012)	442	0.61	[0.46;0.76]
Castro et al. (2014)	26	0.96	[0.31;1.61]
Fuchs et al. (2006)	33	1.01	[0.43;1.59]
Hassler Hallstedt et al. (2018)	127	0.54	[0.32;0.76]
Käser et al. (2013)	41	0.69	[0.28;1.09]
Leh & Jitendra (2013)	25	-0.79	[-1.44;-0.14]
Mohd Syah et al. (2016)	50	1.05	[0.58;1.53]
Nelson et al. (2013)	53	0.65	[0.22;1.09]
Salminen et al. (2015)	21	0.54	[-0.07;1.15]
Stultz (2013)	58	-0.46	[-0.86;-0.05]
RE Model	982	0.55	[0.19;0.9]

Figure 3.1: Forest plot by S. Benavides-Varela et al., 2020, Computers and Education, 157, p.8.

In the rest of this chapter, we will discuss some of the most important aspects of the meta-analysis, as the reviewer's decision on each of them can serve as a researcher degree of freedom and make a turn in the garden of forking paths.

3.2 Effect Size

Effect size is exactly the information the researchers want to reflect in their study (Lakens, 2014). In the context of meta-analysis, effect size plays a crucial role to the extent to be considered "the unit of currency" (Borenstein et al., 2009). Regardless of its importance, many scholars have controversy in how to define this statistic (see Kelley & Preacher, 2012). The definition we use here is "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley & Preacher, 2012, p. 140). This definition is useful as it covers both frequentist and Bayesian frameworks as well as different kinds of studies, be they experimental, correlational, or epidemiological. Effect sizes can be generally categorized into two groups, *standardized* and *unstandardized.* Standardized effect size is a scaled measure that has taken into account the variability of the sample or population of interest (Baguley, 2009). On the other hand, the unstandardized effect has the same unit as the measurement and is affected by the variability of the sample (Baguley, 2009). In some fields, such as psychology, it is more common to use standardized measures for effect size, especially for meta-analysis. The reason is that usually studies use different instruments to measure the same phenomenon (such as different psychological tests), therefore, the comparison of these scales would not be meaningful when we use unstandardized effects (Borenstein et al., 2009). In any case, regardless of the type, effect size usually comes within the bounds of the confidence interval, representing the precision of each study in estimating this value (Borenstein et al., 2009).

To summarize the effect sizes of several studies in a meta-analysis, one should consider several criteria. Firstly, the effect sizes of different studies should comparably measure the same thing and has the same meaning across all studies; Secondly, one should be able to estimate the effect size based on the reported information to have a similar metric of effect size for all of them; In addition, effect sizes should be reliable in terms of having a known distribution so one can compute variance and confidence interval of them, and not leading to error or bias. Lastly, the chosen effect size should be interpretable for our specific research question (Borenstein et al., 2009; Harrer et al., 2021).

One of the most commonly used forms of effect size is based on means. Among those Cohen's d is of great importance as it is one of the most popular effect sizes in meta-analytic investigations (McGrath & Meyer, 2006). Under the assumptions of normality and homogeneity of variance, it captures the mean difference between two groups, which is standardized by within-group pooled standard deviation (Cohen, 1988). The standardized mean difference for a sample is computed as follows:

$$d = \frac{\bar{m}_1 - \bar{m}_2}{SD_{pooled}} \tag{3.1}$$

where $\bar{m}_1$ and $\bar{m}_2$ represent the mean of two groups, and SD_{pooled} represents pooled standard deviation which equals to:

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1 - (n_2 - 1)SD_2}{n_1 + n_2 - 2}} \tag{3.2}$$

where n_1 and n_2 indicate sample sizes for each group, and SD_1 and SD_2 refer to standard deviation of respected groups.

One reason that might contribute to the popularity of Cohen's d is that benchmarks for interpretation of this effect size are widely accepted (McGrath & Meyer, 2006). Cohen (1988) proposed $d = .2$, $d = .5$, and $d = .8$ as small, medium, and large effect sizes respectively in the field of social sciences. Based on his proposal, the effect size of $d = .3$ shows differences that are difficult to detect (e.g., height difference of 15- and 16-year-old girls). Expectedly, the effect size of $d = .8$ refers to completely obvious differences, such as height difference between 13- and 18- year-old girls. The medium effect of $d = .5$ though was defined as "large enough to be visible to the naked eye" (p.26). Although he emphasized 'the terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation' (Cohen, 1988, p. 25), these benchmarks were adopted as heuristics in different fields. It also seems to encourage the heuristic of dismissing "small" effects as unimportant although they can matter under some conditions in fields like psychology (Anvari et al., 2022). Therefore, it is important to consider these heuristics before using the benchmarks to interpret this effect size.

However, Cohen's d does not seem to be the best choice when we are faced with more complex designs such as pretest-posttest-control (PPC) design. PPC is among the most common designs to evaluate the efficacy of programs, such as evaluating an intervention (Morris, 2008). This design is beneficial as it can control for pre-existing differences between treatment and control groups (Morris, 2008). Without a specific method to compute the effect size for this design, one should either use several t-tests, mixed effect analysis, or analysis of covariance (Morris, 2008). However, the question of the precision of these methods was a matter of controversy among scholars. As a result, Morris (2008) suggested d_{ppc2} among other effect size indexes as it comparatively is more favourable in terms of precision, robustness to the heterogeneity of variance, and control for bias. We also use this effect size for the real study application of this study (see chapter 5). Assuming the population variances are homogeneous, d_{ppc2} is computed as follows:

$$d_{ppc2} = C_p[\frac{(M_{post,T} - M_{pre,C}) - (M_{post,C} - M_{pre,T})}{SD_{pooled,pre}}] \tag{3.3}$$

where $M_{pre,T}$ and $M_{post,T}$ are pre and post mean scores for the treatment group and $M_{pre,T}$ and $M_{post,T}$ represent pre and post mean scores for the control group. $SD_{pooled,pre}$ is the pooled standard deviation of the pretest, which only considers pretest standard deviations and sample sizes (n_T and n_C) of both the experimental and control groups,

$$SD_{pooled,pre} = \sqrt{\frac{(n_T - 1)SD^2_{pre,T} + (n_C - 1)SD^2_{pre,C}}{n_T + n_C - 2}} \tag{3.4}$$

and C_p is bias adjustment which is computed as

$$C_p = 1 - \frac{3}{4(n_T + n_C - 2) - 1} \tag{3.5}$$

In PPC design, considering the correlation between pre- and post-test scores is of great importance as it affects the precision of the estimate (Morris, 2008). Equation (3.6) represent the role of this correlation in computing the variance of the d_{ppc2}:

$$\sigma^2(d_{ppc2}) = 2(c_p^2)(1-\rho)(\frac{n_T+n_C}{n_T n_C})(\frac{n_T+n_C-2}{n_T+n_C-4})(1+\frac{\Delta^2}{2(1-\rho)(\frac{n_T+n_C}{n_T n_C})}) - \Delta^2 \quad (3.6)$$

where Δ is the population effect size, and ρ indicates pre-post test correlation. As it is evident, a lower correlation allows greater variance in the distribution which results in lower assigned weight to the study and more conservative results (Benavides-Varela et al., 2020).

Unfortunately, many researchers do not report ρ in their papers, and the meta-reviewer would be left to *guess* the pre-post test correlation. This would open multiple paths in the garden of forking paths, as there might be several plausible values for the correlation. Of course, if the data for each study is publicly available, the meta-analyst can compute the correlation. However, it is barely the case. As an example, in a recent meta-analysis performed by Benavides-Varela et al. (2020), non of the included studies made their data available, and only one them reported the correlation, which left the meta-reviewers to explore different plausible values for the correlation.

Choosing the proper effect size is one of the most crucial steps in performing a meta-analysis, and it is important to choose the most precise and the most informative one for the study. As effect size estimates based on binary data and correlations were beyond the scope of the present study, we did not cover them in this section. More information can be found in Borenstein et al. (2009).

3.3 Fixed-Effect vs. Random-Effects Models

Another important decision in performing meta-analysis is choosing a suitable meta-analytical statistical model for the review. Among these models, the *fixed-effect model* and *random-effects model* are used more frequently (Borenstein et al., 2009). In this section, we will discuss each model and compare their effects on the result of the meta-analysis.

Fixed-effect model (FE) is based on the assumption that there is one true effect size estimate for the population (Borenstein et al., 2009). This means the observed effect of a single study is part of a distribution with a mean θ representing the true effect. The higher the sampling error, the further the observed effect size (T_n) is from the true value (Figure 3.2). Based on this model, if each study had an infinite number of participants, all presented the same effect size estimates (θ).

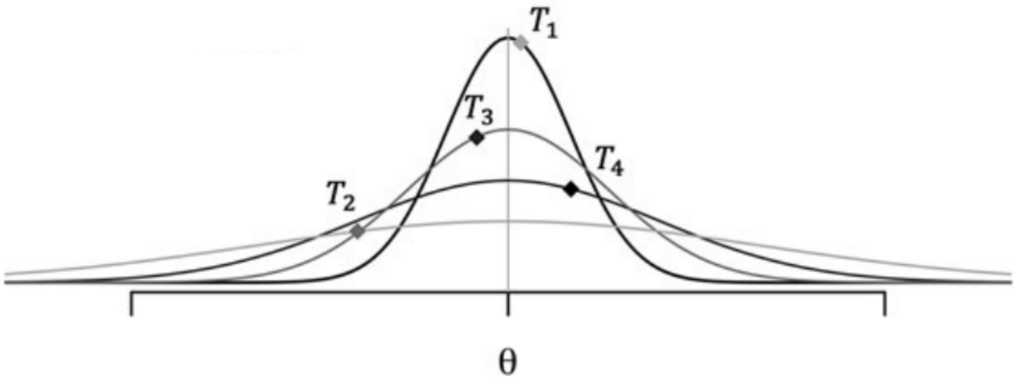


Figure 3.2: The distribution of sampling error in fixed-effect model. Reprinted from Meta-Research: Methods and Protocols (p.45), by S. Kanters, 2022, Humana, New York.

On the contrary, the random-effects model (RE) assumes that the true effects can vary from one study to the other as there is heterogeneity among studies (Borenstein et al., 2009). For example, the efficacy of a treatment may change based on participants' age or gender, or there may be covariates in each study that we did not control for, but they are contributing to the variance. According to this model, if all single studies have an infinite sample size, the effect estimates for all studies would form a normal distribution (Figure 3.3), with the mean of μ representing the average effect (the black triangular) among all possible effect sizes.

As these models have different assumptions, choosing between them can potentially affect the results of the meta-analysis (Dettori et al., 2022), so how we should decide

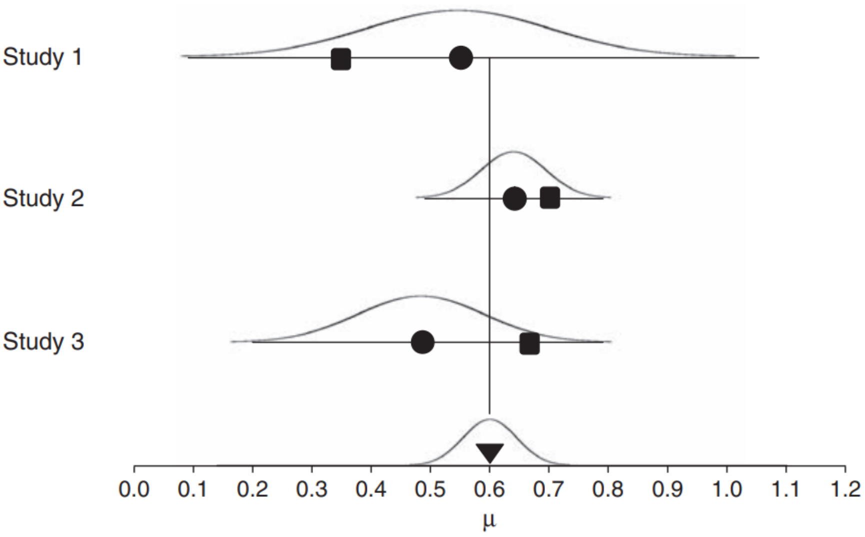


Figure 3.3: Within- and between-study variance in random-effects model. Reprinted from Introduction to Meta-Analysis (p.72), by M. Borenstein et al., 2009, John Wiley and Sons.

which model to use? Many believe the random-effects model is a more appropriate choice (Cai & Fan, 2020; Spineli & Pandis, 2020). This is because the results when using such a model can potentially be generalized to a wider population as this model accounts for between studies diversity as well (Spineli & Pandis, 2020). Moreover, it presents similar results to the FE model in the case of homogeneity among studies (Spineli & Pandis, 2020). However, this viewpoint may lead to overlooking the benefits of FE when it can be used. As we mentioned earlier, meta-analysis can weight single studies according to their precision. This weight is assigned based on the inverse overall variance of each study (i.e., $1/V_{Y_i}$). Because of the basic assumption in the FE model, all variances come from within each study, whereas variance in the RE model, has two contributors (Dettori et al., 2022). As a result, less precise studies get more weight when one uses the random-effects model. Furthermore, if the population effect size is a single value rather than a variety, the FE model can control the Type 1 error rate better (Cai & Fan, 2020). Additionally, when there are few studies involved in the meta-analysis, the effect size estimate of RE can be misleading. As we do not have an accurate estimate of between-studies variance, the summary effect presented by RE can be far off, however, FE can summarize the effects without making inferences about the wider population (Borenstein et al., 2009).

Referring again to the quote at the beginning of this chapter, there is no way to make a certain decision on the underlying model. This choice can always bring some levels of degree of freedom, and it is best to be transparent about the reason behind choosing one analytical model over the others.

3.4 Heterogeniety

As we mentioned at the beginning of this chapter, meta-analysis is not only a tool to summarize the result but also to investigate the patterns among effects. We also saw how this heterogeneity among studies plays a key role in choosing the meta-analytical model. In this section, we will explain this phenomenon in more detail and discuss how to measure it.

By definition, *heterogeneity* is variation in the *true* effect sizes (Borenstein et al., 2009; Higgins, 2008). It can be depicted graphically or can be calculated statistically. Subjectively, one can estimate whether there is variation in true effect size just by looking at the forest plot. As You can see in figure 3.4, B shows narrower confidence intervals for each study than A, indicating they are more accurate. As these accurate studies do not overlap much, there is a higher chance that studies involved in B are truly heterogeneous and do not share a common effect.

However, to objectively assess heterogeneity in meta-analysis, we need to use statistics that are sensitive to heterogeneity and not to the overall variability (Borenstein et al., 2009). Several significance tests have been proposed in the literature to test the presence of heterogeneity (Viechtbauer, 2007). Among them, the Q test is one of the most commonly used ones. Q statistic is a standardized measure computed by summing

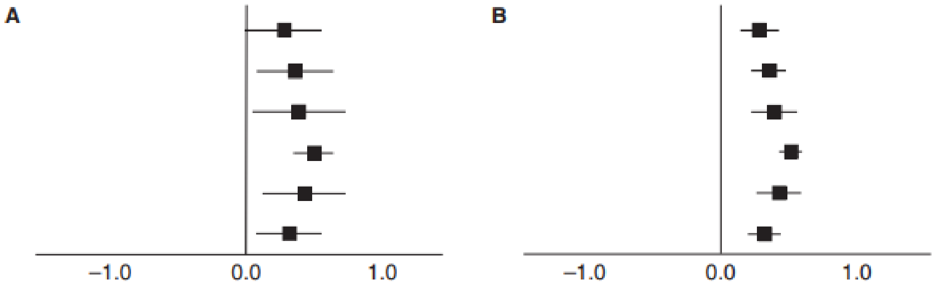


Figure 3.4: Visual representation of heterogeneity. Reprinted from Introduction to Meta-Analysis (p.108), by M. Borenstein et al., 2009, John Wiley and Sons.

the squared deviation of each study's estimate of effect size (Y_i) from the summary effect size (M) while weighting the contribution of each study (Equation (3.7)). Then, by comparing it to the expected value of dispersion when all variability is due to sampling error (degrees of freedom (df)), one can test whether the presence of homogeneity can be statistically rejected (see Borenstein et al., 2009). Reporting it in a confidence interval is also beneficial as it not only shows the precision of the estimate but also gives all information about the significance test (Viechtbauer, 2007).

$$Q = \sum_{i=1}^{k} W_i(Y_i - M)^2 \tag{3.7}$$

Although reporting Q is common in many meta-analyses, we should note that like other significance tests, we should consider some facts about it. Firstly, this test can reject the presence of homogeneity, but it can never be evidence to *accept* it (Baker et al., 2009; Borenstein et al., 2009). It can also be subjected to Type 1 and Type 2 error rates, especially, in case few studies are involved in the meta-analysis and/or high within-study variation (Baker et al., 2009; Borenstein et al., 2009). Moreover, this test can never give information on the magnitude of the heterogeneity and only gives information on its presence (Borenstein et al., 2009).

Another way to quantify heterogeneity in meta-analysis is to estimate τ^2. τ^2 refers to the variance of true effect size, and ideally, it is the value that one adds to within-study variability to compute total variance under the assumption of random-effects model (Borenstein et al., 2009). However, as we never know the exact value of τ^2, we can only estimate it from the observed variability (Borenstein et al., 2009). This parameter and its standard deviation (τ) are in the same metric (one in squared) as effect size. This indicates that one cannot compare these values from different meta-analyses that use different effect size indexes (Borenstein et al., 2009; Huedo-Medina et al., 2006). In addition, we have to take into account that any decision based on τ^2 is subjected to the amount of error we committed in estimating the effect size and the true heterogeneity (Borenstein et al., 2009). Therefore, it is also helpful to report these parameters in a confidence interval to show our precision.

To overcome the shortcomings of previous methods, another statistic was introduced. I^2 aims to measure the true extent of heterogeneity by reflecting the proportion of true difference from the observed variation (Borenstein et al., 2009; Huedo-Medina et al., 2006). In other words, it reflects how much the confidence intervals of the effect size estimates from different studies overlap one another. Therefore, it is free from the estimation and distribution of the true effect size (Borenstein et al., 2009). This statistic can range from 0% representing all variation is caused by sampling error to 100% representing all variation can be explained by true heterogeneity (Huedo-Medina et al., 2006). To facilitate the interpretation of I^2, three benchmarks were proposed, $I^2 = 20$, $I^2 = 50$, and $I^2 = 75$ representing low, medium, and high heterogeneity, respectively (Huedo-Medina et al., 2006). Although τ^2 and I^2 are closely related, τ^2 has the advantage of not being dependent on the number of studies and their precision, while I^2 tends to 100% solely because studies have larger sample sizes (Rücker et al., 2008). On the other hand, I^2 has the advantage of being comparable between different meta-analyses as well as being easier to interpret than τ^2 (Harrer et al., 2021; Huedo-Medina et al., 2006).

As we mentioned, each of these methods to detect and measure the magnitude of heterogeneity has advantages and disadvantages, and using one over the other may result in losing information and misinterpretation of heterogeneity. Although some believe that very high heterogeneity can mean that the studies has nothing in common, therefore, the meta-analysis result would be meaningless (Harrer et al., 2021), others believe that there is no "acceptable" degree of heterogeneity to perform a meta-analysis, and one can accept any degree as long as the eligibility criteria are reasonable and the data would be correct (Higgins, 2008). Certainly, a strong theory can explain the high heterogeneity among studies, however, in absence of a comprehensive theory, we should make sure high heterogeneity is not a case of the apples and oranges.

3.5 Publication Bias

In previous chapters, we discussed a phenomenon that the probability of getting published is affected by the results of the study. It is evident that it will culminate in omitting some evidence from the future meta-analysis selectively (Harrer et al., 2021). Earlier in this chapter, we briefly discussed the file-drawer problem, which is usually used synonymously with publication bias, and explained how meta-analytical results will be distorted by the input. This section will be dedicated to detecting, assessing, and correcting this phenomenon as it has a huge impact on the validity of the meta-analysis. *Publication bias* theoretically refers to "a tendency toward preparation, submission, and publication of research findings based on the nature and direction of the research results" (Dickersin, 2005, p. 13). The concerns about this issue are raised as meta-analyses gain popularity in policy settings (Dickersin, 2005). To depict the severity of the impact of publication bias, imagine implying a certain intervention based on positive outcomes of a meta-analysis of published studies while more unpublished studies present a neutral or negative impact of the same intervention. Not only this would put the lives of patients in danger, but it could also redirect funds from other interventions that may potentially improve the condition of those in need.

However, we should note that it is not the only bias a meta-reviewer faces. Even when negative or non-significant results do get published, they are less likely to be cited and, therefore, harder to be detected, which is referred to as *citation bias* (Harrer et al., 2021). In another case, called *time-lag bias*, studies with faviourable results get published faster (Harrer et al., 2021). Many faviourable studies are also published in more than one journal to expand the impact of the findings, which is called *duplicate publication bias* (Fairfield et al., 2017). There is also *language bias* which refers to overlooking studies that are not in English, the dominant language of publication. This bias can have a more severe impact when the results of studies with two different languages present a contradiction (Harrer et al., 2021). Another source of bias is questionable practices that we covered in Chapter 1 which may hide or alter the results of the study and create distortion in the results of the meta-analysis.

One way to detect publication bias in meta-analysis is by funnel plot. The funnel plot is a simple scatter plot that compares the estimated effect size and the precision of each study (Sterne et al., 2005). As we expect, effect size estimates from studies with higher sample sizes are closer to the population effect size. As a result, when we plot all studies on a certain subject based on their precision and effect size estimates we would get a symmetrical funnel-shaped plot (Sterne et al., 2005). Unfortunately, it is not the case in most fields' literature. In figure 3.5, we can see how publication bias removes part of the funnel and creates an asymmetric plot.

Just by looking at the funnel plot, one can have a subjective understanding of the presence of publication bias. There are also quantitative ways (e.g., Egger's test) to measure how much asymmetry is statistically significant (Borenstein et al., 2009), however, they are beyond the scope of the current study. It is also worth mentioning that publication bias is not the only contributor to asymmetry in the funnel plot. All four aforementioned biases can also make the funnel plot asymmetrical. True heterogeneity

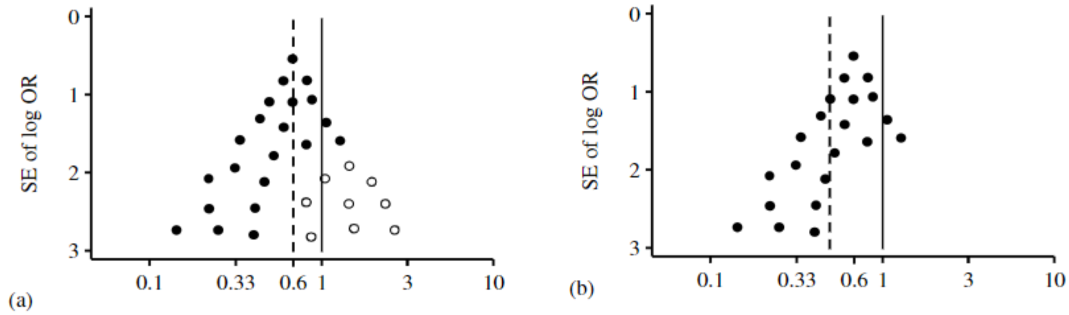


Figure 3.5: Hypothetical example of funnel plot: (a) symmetrical plot in absence of bias (open circles represent not significant findings); (b) asymmetrical plot in presence of publication bias. Reprinted from Publication Bias in Meta-Analysis (p.76), by J. A. C. Sterne et al., 2005, in In Publication Bias in Meta-Analysis (eds H.R. Rothstein et al.), John Wiley and Sons.

(section 3.4) can also change the symmetrical funnel shape of the plot. Moreover, fraud, inadequate analysis, and poor methodological design of the studies can also contribute to asymmetry in the funnel plot (Sterne et al., 2005). Therefore, it is important to consider all contributing factors before using the subsequent methods.

The impact of publication bias on the result of the meta-analysis can be addressed to some extent by some approaches. One common way to correct the bias in meta-analysis is a method called trim and fill. The *trim and fill* method involves a funnel plot to detect publication bias, and then it removes some studies from the asymmetric plot (trimming) until the plot would not be statistically asymmetric. Then, it fills the deleted observations on the opposite side they were initially (Carter et al., 2019). Although this method can correct bias in some cases, there is multiple evidence suggesting that it does not completely correct misestimation of the effect (Carter et al., 2019; Haaf, 2020).

Another approach is to use a group of meta-regression methods. The *precision-effect test* (PET) is a method that assumes a linear relationship between the effect size and standard error and tries to fit them into a linear regression model (Haaf, 2020). However, when the true effect size is not zero, or there would be a non-linear relationship between effect size and standard error this test does not perform as accurately (Carter et al., 2019; Haaf, 2020). As an alternative, one can use the *precision-effect estimate with standard error* (PEESE) method. This approach which is closely related to the previous one assumes a quadratic relationship between effect size and standard error (Haaf, 2020). The logic behind this method is that when there is a true effect, low-powered studies only get published when they overestimate the effect; thus, publication bias is stronger when the standard error is larger (Carter et al., 2019). Simulations also depicted that this method outperforms the other one when there is a true effect (Carter et al., 2019; Haaf, 2020). As a result, it is recommended to use a mixture of these methods called *PET-PEESE* (Figure 3.6). If the PET estimate is statistically significant, it is recommended to use PEESE and reverse (Carter et al., 2019; Haaf, 2020). Although many simulations showed that the performance of PET-PEESE was promising, it seemed to have a poor performance when there is high heterogeneity and when there are only a few low-sample-sized studies available for meta-analysis (Carter et al., 2019).

There are other approaches to address publication bias, which we did not cover in this section. However, it is important to note each method has different advantages and disadvantages, and ways to compare these methods also have certain drawbacks (Haaf, 2020). One way to immune the estimate of the meta-analysis from publication bias is to perform many-labs studies (Haaf, 2020). Based on this approach, an identical procedure is followed by several research teams, and a meta-analysis is performed to summarize the results of all of them (Ebersole et al., 2016). This method can also work as a reference point to compare the results of bias-corrected meta-analyses on a similar topic to check which method performs better than the others (Haaf, 2020). In the end, we should accept that no matter which correction approach we choose, we can never free the result of a meta-analysis of the literature from bias.

From choosing which studies to include in the meta-analysis to choosing the effect size

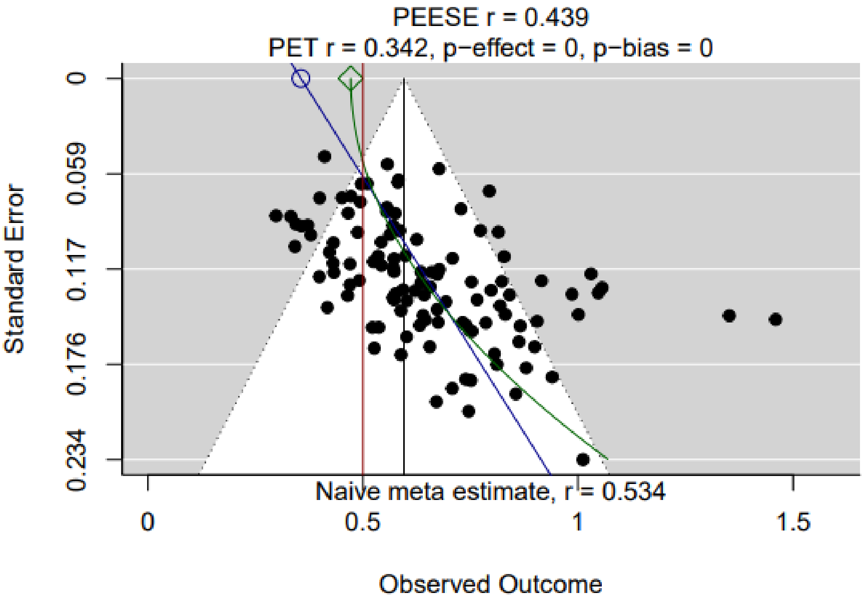


Figure 3.6: PET-PEESE anlysis. As the PET analysis is significant, PEESE should be used. You can see both are underestimating the true effect (red line) but PEESE has more accurate estimate than PET in this case. By J. Haaf, 2020, PsyArXiv [Preprint], p.9.

index, managing the correlations, handling the heterogeneity, choosing meta-analytical models, and the method to correct the publication bias, a researcher faces options that can affect the result of the meta-analysis. Although they are necessary decisions to take, one should be aware of their impact. As we discussed in the previous chapter, multiverse meta-analysis can present a solution. However, even considering few researcher degrees of freedom will culminate in an explosion of results for the multiverse. Therefore, a framework to have an informed interpretation of the overwhelmingly large results of multiverse meta-analysis is necessary. In subsequent chapter, we will discuss our proposed framework and implement it on a real case study in chapter 5.

Chapter 4

Proposed Framework

In this chapter, we will discuss different aspects of our proposal on how to summarize the result of a multiverse meta-analysis through tabular and graphical representations alongside assessing the impact of different choices on the final results. Our proposed framework consists of three main facets, summary tables, graphical representations, and analysis of variance, which will be discussed in more detail in this chapter.

4.1 Summary Tables

The first step in all meta-analyses, including multiverse meta-analyses, should be to present some essential information about the included studies. This step not only serves the purpose of transparency but also helps the audience to better understand the multiversal approach that will be implemented. This table should involve information such as the number of outcomes and measurements used in each study, the number of participants in control and experimental groups, and effect sizes and standard error reported by each study.

Moreover, we need to present a table to numerically show the meta-analytical results for each combination in the multiverse of choices. This information can easily be extracted from the multiverse matrix. In the multiverse matrix, each row represents a combination of choices and the result of the meta-analysis performed using those combinations. This table involves information on the estimated effect within its confidence bounds for each meta-analysis, along with the estimated error and τ^2 estimation for those meta-analyses that used the random-effects model.

Finally, we propose to present a table with relevant descriptive statistical indices (i.e., minimum, maximum, quartiles, median, mean, and standard deviation) to summarize the distribution of all estimated effects sizes calculated in the multiverse meta-analysis.

4.2 Graphical representations

Visualization is a key way to interpret and communicate the result of analysis (Allen et al., 2021). Therefore, it is crucial to use plots that are informative and yet not too complex to understand.

We propose to use the raincloud plot to show the overall distribution of effect sizes according to the plausible choices used in the multiverse (e.g., the distribution of effect sizes according to the meta-analytic model, the distribution of effect sizes according to the meta-analytic model and the imputed pre-post test correlation, . . .).

Raincloud plot is a form of visual representation that in many ways exceeds its predecessors like barplot, dot plot, or even the most recent violin plot (Allen et al., 2021). A raincloud plot consists of a split-half of a violin (as the original violin plot duplicates information by mirroring the other half), raw data points, and a boxplot to visualize the quartiles of the distribution (Allen et al., 2021). In Figure 4.1, you can see an example of a raincloud plot.

Raincloud plots will also be applied to show the overall distribution of the standard errors of the estimated effect sizes conditional on the plausible choices. In case of random-effects models, a scaterplot of τ^2 estimates will be presented as well.

Finally, a bar graph will be used to highlight the contribution of each plausible choice

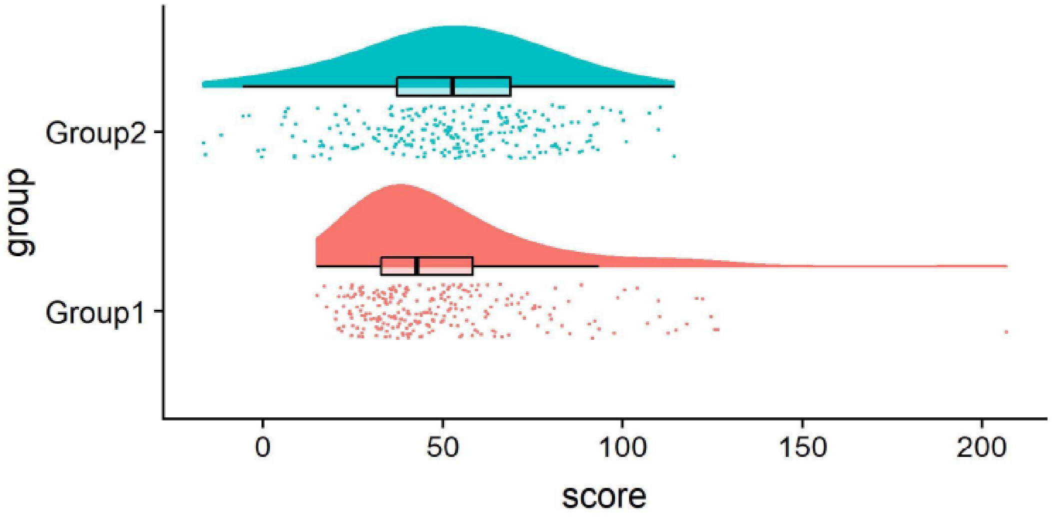


Figure 4.1: An example of a raincloud plot. Reprinted from Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved], by M. Allen et al., 2021, Wellcome Open Research, 4(61), p.5.

in explaining the variability of the effect sizes obtained in the multiverse analysis (see also next section).

4.3 Analysis of Variance

When factors with two or more levels are involved in the study, a researcher usually is interested to check whether membership in each level of the variable explains the variation in the outcome of the study. The analysis of variance (ANOVA) is a common and popular way to check for such contributions. The idea of multiverse analysis is to consider several levels for each variable; Therefore, using ANOVA is justified and informative.

To run the analysis of variance, we propose to simulate data based on the estimated effect and standard error of each meta-analysis to reproduce the sampling distribution of each combination. In this way, we simulate true population-level estimates for each combination, and thus, we make a more robust conclusion from ANOVA. Moreover, we propose to use contrast coding instead of dummy coding for levels of each factor to have subtle and independent comparisons between multiple levels of factors.

Specifically, we propose to report eta-squared, η^2, (a measure of effect size) for each variable involved in the analysis of variance (see Richardson, 2011). η^2 measures the proportion of variation in the outcome variable (in this case, the simulation based on estimated effects) that is associated with being assigned to different levels of a categorical variable (Richardson, 2011). Using this statistic, we can realize how much each variable explains the overall variance in the outcome.

In the next chapter, we will run a multiverse meta-analysis on a real dataset and implement this framework on the results.

Chapter 5

Case Study

In this chapter, we first discuss the data frame we used and later discuss the result of implementing our framework on this dataset.

5.1 Dataset

For our study, we used data from a meta-analysis performed by Daros et al. (2021), which was accessible from the OSF repository for the study (https://osf.io/56fvu). This meta-analysis aims to investigate whether improvements in emotion regulation (ER) skills and emotion dysregulation are associated with improvements in anxiety and depression symptoms in psychological treatment targeting the latter two for patients aged 14-24. They included 88 peer-reviewed studies that were written in English, reporting 90 randomized control trials (RCT) that measured depression and/or anxiety and emotion regulation as an outcome for meta-analysis. They also separately analyzed 55 non-RCT studies that met the inclusion criteria and used them for comparative analysis with the primary findings of the meta-analysis. The results of the multivariate random-effects meta-analysis showed that psychological treatment alleviated anxiety, depression, disengagement ER skills and emotion dysregulation and elevated engagement ER skills.

Although they did not explicitly mention the effect size measure they used for their analysis, trying the Morris formula drew close results to the reported effect sizes, suggesting they probably calculated d_{ppc2} with Hedges' correction for each study. However, the magnitude of the pre-post test correlation they used was not clear. It was also not clear what correlation they used to aggregate the different measurements of the same construct. The only correlation they reported was $\rho = 0.70$ for correlation among variables.

Although nearly all studies included in this meta-analysis were randomized control trials, not all of them had similar designs. Some of the studies had two experimental groups and one control group or two control groups and one experimental group, and in one case, there was no control group. For the current study, we excluded the study without a control group and another study whose sample size was not reported in the dataset. Additionally, we almost randomly chose one treatment/control group in cases with more than one as using one group for two comparisons adds another level of dependency to the design. Of course, it is necessary to have a justification rather than a random choice for excluding information from the dataset; however, our study aims to illustrate a methodological approach and not to investigate a psychological construct; thus, we decided on random selection.

Moreover, the original meta-analysis was a multivariate meta-analysis studying 5 constructs. In our study, we only focused on depression (one of the main constructs in the original study). The reason for this choice was to decrease complexity by removing one source of dependency between statistical units and remaining committed to the educational purpose of this project.

In the current study, we consider plausible options for three arbitrary choices. First one is pre-post test correlation. This correlation is needed to compute variance which is needed for weighting the included studies. The second one is the correlation between multiple measurements of the same construct. This correlation is needed to aggregate results from several measurement methods into one and having one outcome for each study (see below). The last arbitrary choice is related to the meta-analytical model. As mentioned in Chapter 3, meta-analytical models are frameworks to combine and analyze the results of multiple studies.

5.2 Implementing the Framework

All statistical analyses and data manipulation on this section were performed using the R programming language (version 4.2.2), with the following packages: `metafor` (Viechtbauer, 2010) was used to conduct meta-analyses, and `ggplot2` (Wickham, 2016) and `ggrain` (Judd, 2023) were used for visualization.
As we mentioned in the previous chapter, the first step in our framework is to present an overview of essential information about the dataset we are working with.

Table 5.1: Summary of studies included in the multiverse meta-analysis

study	comparisonN1.N2	measure	N1	N2	yi	std_err	vi
Ahmad 2020	FullMBCT-Control	PHQ-9	39	38	0.479	0.229	0.052
Araya 2013	CBT-Control	BDI-II	1219	1289	0.039	0.040	0.002
Auslander 2017	CBT-Control	CDI	17	10	0.103	0.387	0.150
Baert 2010-Study1	ABM-control	BDI-II	25	23	-0.494	0.289	0.083
Bentley 2018	Workshop-AO	DASS-D	68	70	0.338	0.171	0.029
Biggam 2002	PST-control	HADS-D	23	23	0.711	0.299	0.089
Bluth 2016	MCBT-control	SMFQ-dep	16	18	0.515	0.341	0.116
Burckhardt 2018	DBT-control	CES-D	50	46	0.234	0.203	0.041
Chambers 2015	MBCT+TAU-TAU	CES-D	20	21	0.313	0.308	0.095
Chambers 2015	MBCT+TAU-TAU	HAMD	20	21	0.430	0.310	0.096
Clore 2006	FluencyT-TRec	BDI-II	15	15	0.324	0.358	0.128

Table 5.1: Summary of studies included in the multiverse meta-analysis *(continued)*

study	comparisonN1.N2	measure	N1	N2	yi	std_err	vi
DamaioNeto 2020	M-psyched	DASS-D	70	71	0.213	0.168	0.028
Delgado-Pastor 2015	MInteroG-Control	BDI	15	14	0.992	0.384	0.147
Delgado 2010	MSBR-Relax	BDI	15	17	0.169	0.346	0.120
Dereix-Calonge 2019	ACT-WL	DASS-D	43	42	0.823	0.224	0.050
DeVoogd 2016a	DPT-DPplacebo	CDI	128	48	0.003	0.169	0.028
DeVoogd 2016b	EWM-placebo	CDI	129	39	0.189	0.182	0.033
DeVoogd 2017	Pic-control	CDI	44	39	0.248	0.219	0.048
DeVoogd 2018	CBMi-placebo	CDI	134	39	0.039	0.181	0.033
Diaz-Gonzalez 2018	MSBR+TAU-TAU	SCL-anx	41	39	0.257	0.222	0.049
Donaldson 2005	CBTER-support	CES-D	15	16	0.255	0.352	0.124
Dowling 2019	CBTER-control	DASS-d	245	250	0.010	0.090	0.008
Dvorakova 2017	MSBR-WL	phq9	55	54	0.445	0.193	0.037
El Morr 2020	Mindful-WL	PHQ9	79	80	0.439	0.160	0.026
Eskin 2008	PST-WL	BDI	27	19	1.125	0.317	0.100
Falsafi 2016	MindSC-Control	BDI	21	23	0.851	0.310	0.096
Gouda 2016	MSBR-WL	HADS-dep	15	14	-0.287	0.363	0.132
Griffiths 2019	MBT-tau	RCADS-dep	22	26	0.041	0.285	0.081
Gross 2018	ACT-PST(cbt)	CCAPS-dep	11	11	0.022	0.410	0.168
Gu 2018	MBCT-WL	BDI	28	26	0.758	0.278	0.077
Hamdan-Mansour 2009	CBT-AO	BDI	44	40	1.433	0.243	0.059
Haukass 2018	ATT-MSC	PHQ9	40	41	0.022	0.220	0.048
Hetrick 2017	iCBT-TAU	CDRS	26	24	0.230	0.280	0.078
Hetrick 2017	iCBT-TAU	RADS	26	24	0.196	0.279	0.078
Hoorelbeke 2015	CCT-visST	BDI	25	22	0.065	0.288	0.083
Horowitz 2007	CB-Control	CDI	112	169	0.203	0.122	0.015

Table 5.1: Summary of studies included in the multiverse meta-analysis *(continued)*

study	comparisonN1.N2	measure	N1	N2	yi	std_err	vi
Horowitz 2007	CB-Control	CES-D	112	169	0.142	0.122	0.015
Idsoe 2019	CBT-TAU	CES-D	133	95	0.330	0.135	0.018
Keng 2019	Mindfulness-control	DASS-dep	28	29	0.172	0.262	0.069
Ko 2018	MedComp-group	CES-D	18	16	0.248	0.337	0.113
Kowalenko 2005	CBT-AO	CDI	41	41	0.371	0.221	0.049
Kuosmanen 2017	SPARX-Control	SMFQ	30	36	-0.102	0.244	0.060
Levin 2014	webACT-WL	DASS-dep	37	39	0.231	0.228	0.052
Levin 2017	webACT-WL	CCAPS-dep	40	39	0.183	0.223	0.050
Levin 2020	Mindapp-control	CCAPS-dep	10	13	0.945	0.429	0.184
Lindqvist 2020	psydyn-control	MADRS	38	38	0.694	0.234	0.055
Livheim 2015-s1	ACT-tau	RADS-depression	32	26	0.512	0.265	0.070
Livheim 2015-s2	ACT-tau	DASS-dep	15	17	-0.081	0.345	0.119
Maestas 2012	AMTtrad-control	BDI	67	68	0.041	0.171	0.029
McIndoo 2016	BehAct-WL	BDI-ii	16	14	0.764	0.369	0.137
McIndoo 2016	BehAct-WL	HRSD	16	14	1.207	0.389	0.151
Mogoase 2013	ConcreteT-wl	BDI	20	21	0.161	0.307	0.094
Mokrue 2013	CBTskills-WL	BDI	54	30	1.261	0.246	0.060
Morris 2015	CBT(CR)-writing	CES-D	84	82	0.137	0.155	0.024
Muto 2011	ACTwb-WL	DASS-d	35	35	0.209	0.237	0.056
Nguyen-Feng 2015	SM- WL	DASS-d	329	171	-0.140	0.094	0.009
Oldenzki 2020	Mind(hyp)-WL	PDP-dep	14	16	1.169	0.387	0.150
Puskar 2003	TKC-control	RADS	46	43	0.457	0.213	0.045
Rasanen 2016	iACT-control	BDI	33	35	0.829	0.250	0.063
Rasanen 2016	iACT-control	DASS-dep	33	35	0.831	0.250	0.063
Reddy 2013	CBT-waitlist	QIDS	35	35	0.021	0.236	0.056

Table 5.1: Summary of studies included in the multiverse meta-analysis *(continued)*

study	comparisonN1.N2	measure	N1	N2	yi	std_err	vi
Richards 2016	iCBT-WL	BDI-ii	70	67	0.415	0.172	0.030
Rizzo 2018	DS-knowledge	BDI	59	50	0.197	0.191	0.037
Rudd 1996	CBTint-TAU	BDI	143	121	0.276	0.124	0.015
Sheffield 2006	combined-control	CDI	112	149	0.002	0.125	0.016
Sheffield 2006	combined-control	CES-D	112	149	0.149	0.125	0.016
Shomaker 2017	gMindful-cCBT	CES-D	17	16	0.793	0.354	0.125
Singhal 2018	CBTskill-psyed	CDI	65	55	3.087	0.270	0.073
Singhal 2018	CBTskill-psyed	CES-D	65	55	2.422	0.240	0.058
Slee 2008a	CBT-TAU	BDI	42	48	0.923	0.220	0.049
Slee 2008b	CBT-TAU	BDI-II	40	42	0.340	0.220	0.049
Song 2015	MSBR-WL	DASS-d	21	23	0.654	0.304	0.093
Stasiak 2012	cCBT-control	CDRS	17	17	0.769	0.348	0.121
Stasiak 2012	cCBT-control	RADS-2	17	17	0.832	0.350	0.122
Teng 2019	ABM-WL	BDI	30	22	0.120	0.277	0.077
Topper 2017	gRFCBT-wWL	BDI	82	85	0.644	0.158	0.025
Uliazsek 2016	DBT-PPT	SCL-90-dep	27	27	0.604	0.274	0.075
Vrijsen 2018-s1	CBM-placebo	BDI-II	51	50	0.161	0.198	0.039
Vrijsen 2018-s2	CBM-placebo	BDI-ii	46	54	0.052	0.199	0.040
Wimmer 2019	MSBR-passive	HADS-dep	51	38	0.405	0.215	0.046
Yang 2015	ABM-AO	BDIii	27	23	1.146	0.302	0.091
Yang 2016	ABM-placebo	CES-D	23	22	0.095	0.293	0.086
Yang 2016	ABM-placebo	HAMD	23	22	0.514	0.298	0.089
Yusoff 2015	workshop-control	BDI	88	83	0.295	0.153	0.023
Zemestani 2015	MCT-Control	BDI	15	15	5.273	0.768	0.590

Zhang 2019	gMSBR-wl	BDI	28	28	0.617	0.270	0.073

Table 5.1 shows that some studies used more than one measurement method to assess depression. Using different measurements adds another level of dependency as each measurement method is assessing the same construct from the same participants. Therefore, it is essential to address this dependency by aggregating the results of different measurement methods to have one outcome for each study in the meta-analysis. According to Borenstein et al. (2009), in order to aggregate the results from two measurement methods into one outcome, we should consider the correlation between the different methods. In practice, one should look at the literature to assess the correlation between them. However, we used different plausible correlations to depict the impact of this decision when it is taken arbitrarily. In the end, we had three sources of possibilities for our multiverse analysis, the correlation between pre-post test (rmorris), the aggregation correlation between different measurement methods (ragg), and different meta-analytical models. As mentioned in Section 2.1, the mission of multiverse analysis is to consider different **plausible** choices rather than exploring all. Thus, we decided to use a reasonable range of 0.6 to 0.8 for pre-post test correlation, and 0.4 to 0.8 for aggregation correlation with 0.05 increment for both choices. Using the `metafor` package (Viechtbauer, 2010), we performed a meta-analysis for all plausible combinations and stored the results in a matrix. Table 5.2 presents the first 8 rows of our multiverse matrix. The complete version of this table is available in Appendix.

The τ^2 row for the fixed-effect model is empty as this model assumes 0 heterogeneity by definition. As it is visible from this table, there is a big difference between the calculated effect size for fixed and random effects models while keeping other variables constant. Whereas the estimates are closer to one another when keeping the meta-analytical model the same. This difference between the numerical magnitude of the calculated effect size by the two models for the entire matrix is visible from the descriptive information

Table 5.2: The multiverse matrix

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.6	0.40	fit fixed	d_{ppc2}	0.238	0.018	13.061	<0.001	0.202	0.274
0.6	0.40	fit fixed	τ^2	0.000					
0.6	0.40	fit random	d_{ppc2}	0.418	0.056	7.483	<0.001	0.308	0.527
0.6	0.40	fit random	τ^2	0.188	0.039				
0.6	0.45	fit fixed	d_{ppc2}	0.237	0.018	12.987	<0.001	0.202	0.273
0.6	0.45	fit fixed	τ^2	0.000					
0.6	0.45	fit random	d_{ppc2}	0.417	0.056	7.495	<0.001	0.308	0.526
0.6	0.45	fit random	τ^2	0.186	0.038				

Note:

'd_{ppc2}' and 'τ^2' are overall estimates for any single meta-analysis.

Table 5.3: Descriptive statistics for estimated effect size calculated by the multiverse meta-analysis.

	n	mean	sd	median	min	max	range	se	Q1	Q3
Fixed-effect	45	0.228	0.006	0.229	0.216	0.238	0.022	0.001	0.223	0.233
Random-effect	45	0.416	0.002	0.417	0.411	0.420	0.009	0.000	0.415	0.418

presented in Table 5.3.

We can also depict this information by using the raincloud plot. Figure 5.1 shows the difference between the effect size estimates based on different choices for the model and pre-post correlation.

As it is visible, while there is no remarkable difference in estimated effect size using different pre-post test correlations, there is a considerable difference in calculated overall d_{ppc2} when we choose different models. The same pattern also appears when we plot the effect size based on the meta-analytical model and aggregation correlation (Figure 5.2), as well as when we plot the estimated the standard errors for effect sizes based on these decisions (Figure 5.3).

Focusing only on the random-effects model, it is also interesting to explore the impact of different correlation choices on the estimation τ^2. Figure 5.4 depicts this impact. Although we can see choosing a lower correlation for aggregation can culminate in a higher estimate of τ^2, it does not seem to be remarkable. Similarly, this estimate is not

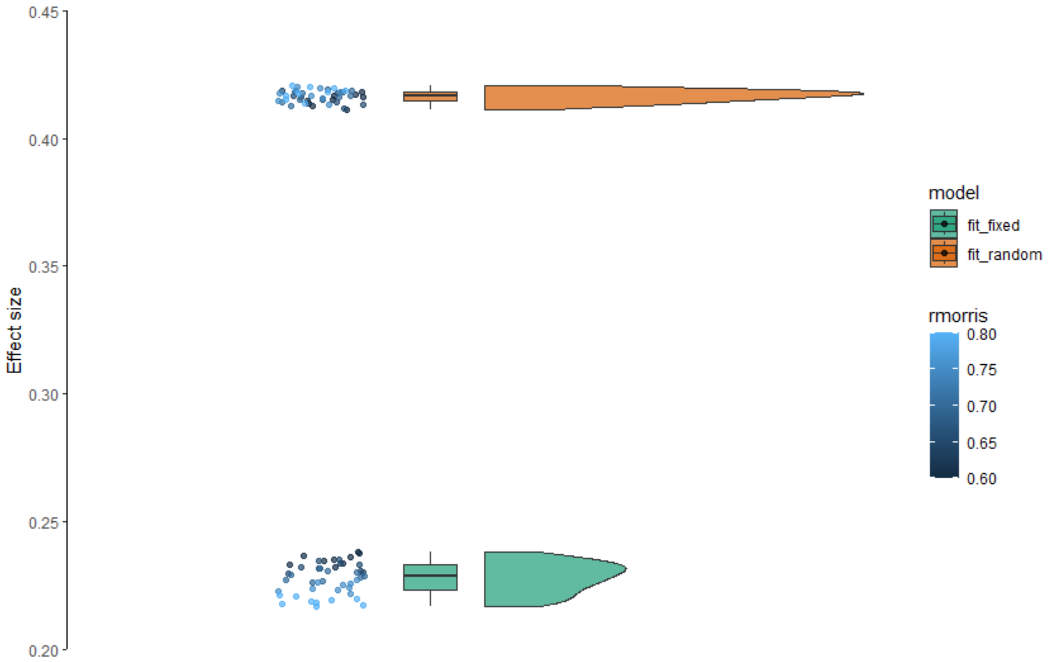


Figure 5.1: The raincloud plot of the overall d_{ppc2} calculated by multiverse meta-analysis according to the meta-analytical model and the correlation between pre-post test. Note that the axes do not start from zero.

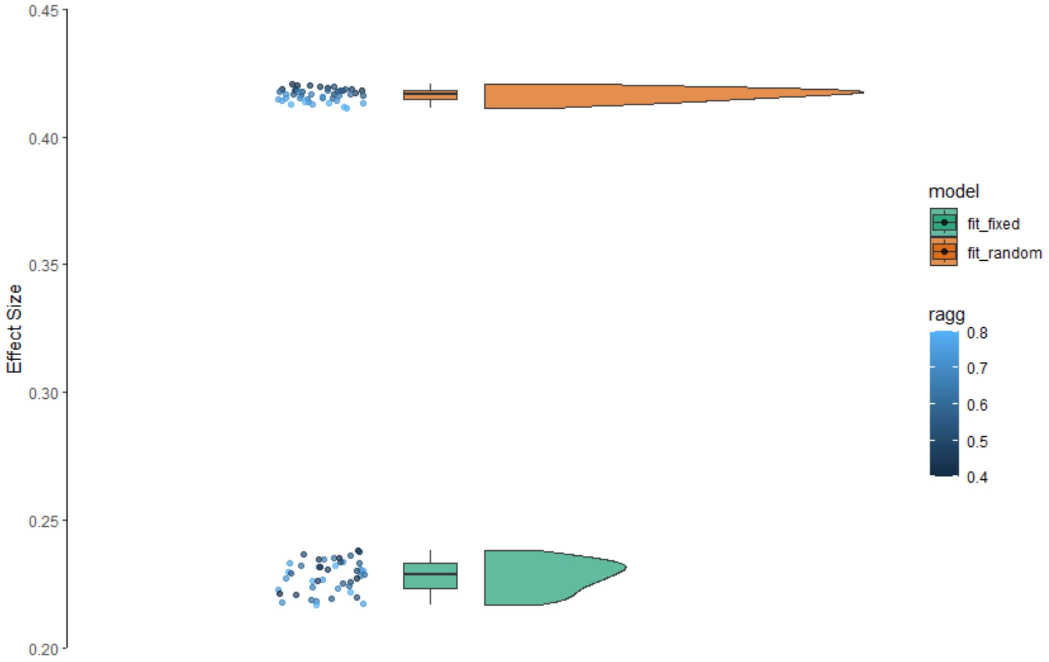


Figure 5.2: The raincloud plot of the overall d_{ppc2} calculated by multiverse meta-analysis according to the meta-analytical model and aggregation correlation. Note that the axes do not start from zero.

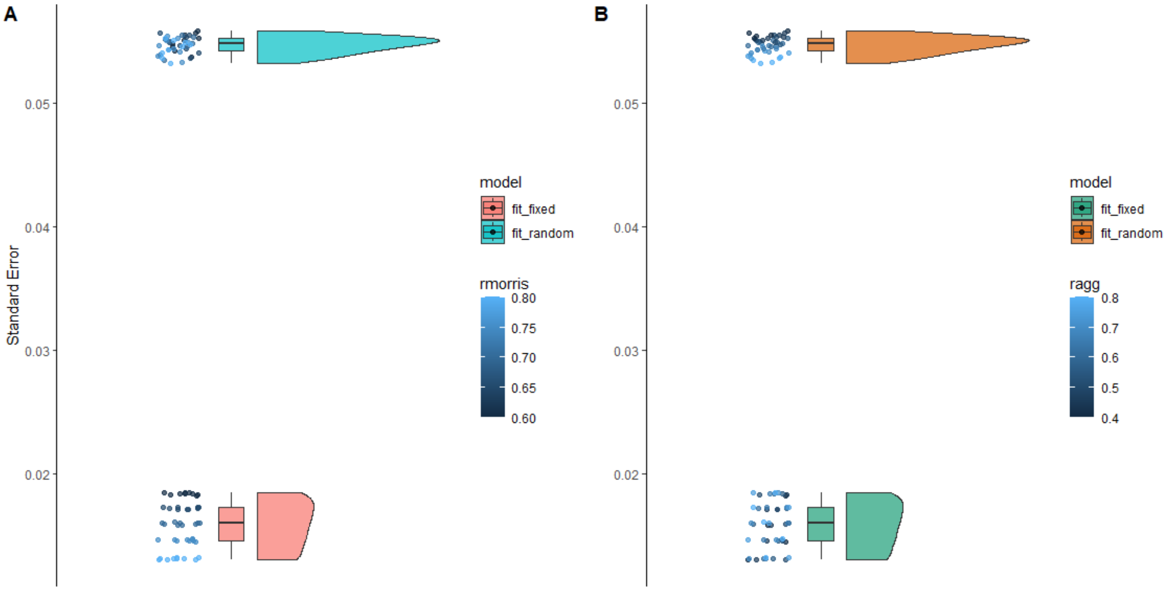


Figure 5.3: A. The raincloud plot of the standard error estimates calculated by multiverse meta-analysis according to the meta-analytical model and pre-post test correlation. B. Similar plot based on the meta-analytical model and aggregation correlation. Note that the axes do not start from zero.

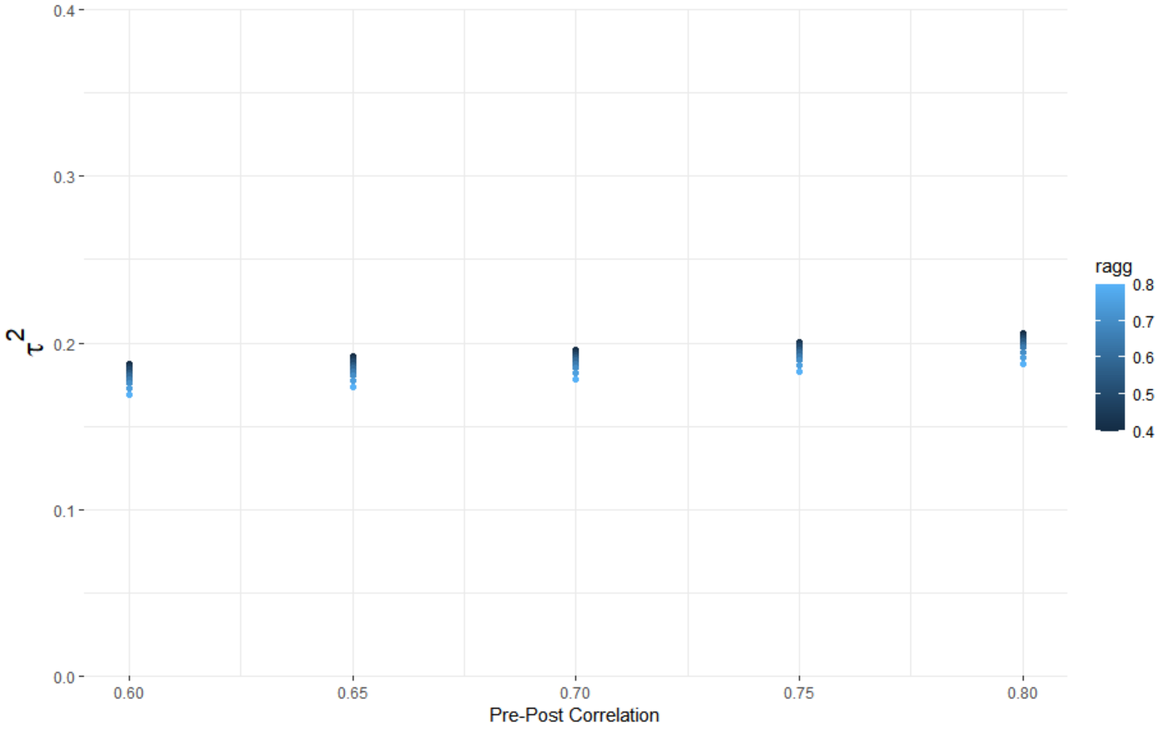


Figure 5.4: The difference between τ^2 estimates for random-effects model according to different choices for the correlation between pre-post test and aggregation correlation. Note that the x asix does not start from zero.

considerably different for choosing different pre-post test correlations.

For the analysis of variance, we used a large simulated dataset. We drawn 10000 data from normal distribution for each combination where the mean is equal to the calculated effect size and *SD* equals to estimated standard error. Given the large number of simulated data ($n_{total} = 900000$), we were interested only on the effect size (η^2). Table 5.4 shows the results of the analysis of variance. In line with the results of previous tables and figures, the meta-analytical model explains the largest variance in our dataset ($\eta^2 = .804$).

Table 5.4: The results of the Analysis of Variance for the arbitury decisions

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η^2
rmorris	4	4.872	1.218	742.677	<0.001	0.004
ragg	8	3.275	0.409	249.597	<0.001	0.002
model	1	8024.844	8024.844	4893021.219	<0.001	0.804
Residuals	899986	1476.030	0.002			

Note:

Total R-squared = 0.81

This results are also graphically depicted by Figure 5.5

5.3 Discussion

In this chapter, we applied our framework on a real dataset. Aligning with the findings of the original study, we found consistent positive effects of different interventions (i.e., CBT, mindfulness, acceptance/ER-based (e.g., DBT), and other therapeutic orientations such as family therapy) on depression across all our estimates.

Regarding the multiverse of choices, the results indicated that the choice of the meta-analytical model has the greatest impact on the results of this particular case, while the rest contributed negligibly to the variability of the outcomes. This means that

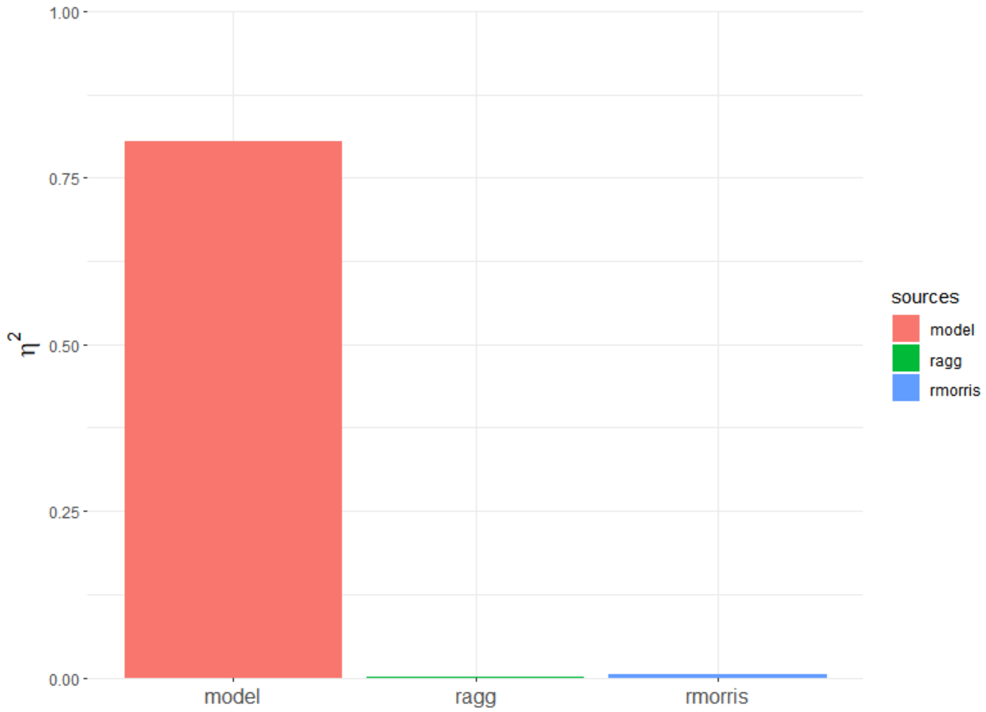


Figure 5.5: Visualization of explained variance based of sources of variability.

choosing different pre-post test correlation and aggregation correlation will not remarkably change the calculated effect size. Therefore, while the choice remains in the range of plausible options, no matter what correlations we choose, our results remain robust. However, making an arbitrary selection of the meta-analytical model can lead to misestimated results. Although it is not common to choose a fixed-effect model for a meta-analysis with so many studies like this one, adopting both models serves the educational purpose of this research. Moreover, for many meta-analyses, especially in the field of psychology, the model choice remains an arbitrary decision due to the few numbers of studies involved.

All in all, these results demonstrate the importance of adopting a multiverse approach for meta-analyses to have more robust and informative results.

Chapter 6

Conclusions

Psychology is currently facing a significant credibility crisis, and restoring trust in the field requires a commitment to openness and transparency. Multiverse analysis provides a valuable approach to enhance transparency by considering all plausible choices that a researcher could take. Given the impact of meta-analytical studies and the great number of arbitrary decisions involved in them, they particularly benefit from an approach that ensures the robustness of their results. Therefore, adopting the multiverse framework is well justified for meta-analyses.

In this regard, the present study aimed to propose a framework to summarize the overwhelmingly large results of a multiverse meta-analysis. In addition, we sought to investigate the impact of different choices on the variability of the results calculated by the meta-analysis.

While our proposed framework is a significant step forward, we acknowledge that it has certain limitations that need further attention and refinement.

First, our framework does not address other researchers' degrees of freedom associated with meta-analysis. Other levels of dependency are involved with more complex meta-analytical designs, such as handling different variables in multivariate meta-analysis or tackling moderators in meta-regression. Another uninvestigated degree of freedom is regarding the bias correction method. Although bias correction is crucial to have accurate and reliable results in a meta-analysis, our framework did not explore it. Moreover, although addressing influential studies is part of the usual flow of meta-analysis, our frameworks did not consider it. Examining the effect of considering/not considering the influential studies on the result of multiverse meta-analysis will improve the robustness and informativeness of the results.

Second, this study did not explore outlier detection in relation to the combinations involved in the multiverse. In theory, it is plausible that certain combinations would yield effect sizes that are extremely different from others. Presenting this information provides valuable insights into the robustness and reliability of calculated effect sizes across different combinations.

Lastly, our framework lacks the elements necessary for the Bayesian approach to multiverse meta-analysis and exclusively considered the frequentist viewpoint. Although this material did not claim to cover both frequentist and Bayesian approaches, the absence of one restricts the comprehensiveness of the current proposal.

The proposed framework provides a transparent approach for presenting and exploring the impact of various choices on the result of the meta-analysis. It also offers a more informative understanding of the robustness and uncertainty of the results and thus improves the credibility and replicability of the meta-analytic research. It is hoped that this framework encourages other researchers to adopt a multiverse approach to further their confidence in the result of the meta-analysis, and improves the communication of the tremendous number of results in such studies.

References

Abendroth, A., Parry, D. A., Roux, D. B. le, & Gundlach, J. (2020). An analysis of problematic media use and technology use addiction scales – what are they actually assessing? In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology* (pp. 211–222). Springer International Publishing. https://doi.org/110.1007/978-3-030-45002-1_18

Aczel, B., Szaszi, B., Nilsonne, G., Van Den Akker, O. R., Albers, C. J., Van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., et al. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, *10*. https://doi.org/10.7554%2FeLife.72185

Akker, O. R. van den, Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., et al. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2625

Allen, M., Poggiali, D., Whitaker, K., Marshall, T., Langen, J. van, & Kievit, R. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*, *4*(63). https://doi.org/10.12688/wellcomeopenres.15191.2

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of Clinical*

Psychiatry, *82*(1), 25941. https://doi.org/10.4088/JCP.20f13804

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2022). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*, 17456916221091565. https://doi.org/10.1177/17456916221091565

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "many analysts, one data set" project. *Socius*, *7*, 23780231211024421. https://doi.org/10.1177/23780231211024421

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., From the Health Outcomes, Policy, & Collaborative Group, E. (HOPE). (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice*, *63*(10), 1426–1434. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Bakker, B. N., & Lelkes, Y. (2018). Selling ourselves short? How abbreviated measures of personality change the way we think about personality and politics. *The Journal of Politics*, *80*(4), 1311–1325. https://doi.org/10.1086/698928

Bakker, M., Veldkamp, C. L., Van Den Akker, O. R., Van Assen, M. A., Crompvoets, 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

Banks, G. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. In *Journal of Management* (No. 1; Vol. 42, pp. 5–20). Sage Publications Sage CA: Los Angeles, CA. https://doi.org/10.1177/0149206315619011

Bartlett, J. E., & Charles, S. J. (2022). Power to the people: A beginner's tutorial to power analysis using jamovi. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.3078

Bartoš, F., & Maier, M. (2022). Power or alpha? The better way of decreasing the false discovery rate. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2460

Benavides-Varela, S., Callegher, C. Z., Fagiolini, B., Leo, I., Altoè, G., & Lucangeli, D. (2020). Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers & Education*, *157*, 103953. https://doi.org/10.1016/j.compedu.2020.103953

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108–113. https://doi.org/10.1002/smj.975

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Cai, Z., & Fan, X. (2020). A comparison of fixed-effects and random-effects models for multivariate meta-analysis using an SEM approach. *Multivariate Behavioral Research*, *55*(6), 839–854. https://doi.org/10.1080/00273171.2019.1689348

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chambers, C. D. (2017). *The seven deadly sins of psychology*. Princeton University Press.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass*, *8*(5), 201–213. https://doi.org/10.1111/spc3.12099

Crocetti, E. (2016). Systematic reviews with meta-analysis: Why, when, and how? *Emerging Adulthood*, *4*(1), 3–18. https://doi.org/10.1177/2167696815617076

Daros, A. R., Haefner, S. A., Asadi, S., Kazi, S., Rodak, T., & Quilty, L. C. (2021). A meta-analysis of emotional regulation outcomes in psychological interventions for youth with depression and anxiety. *Nature Human Behaviour*, *5*(10), 1443–1457. https://doi.org/10.1038/s41562-021-01191-9

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication." *Perspectives on Psychological Science*, 1490–1505. https://doi.org/10.1177/17456916211041116

Dettori, J. R., Norvell, D. C., & Chapman, J. R. (2022). Fixed-effect vs random-effects models for meta-analysis: 3 points to consider. *Global Spine Journal*, *12*(7), 1624–1626. https://doi.org/10.1177/21925682221110527

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In *Publication bias in meta-analysis* (pp. 9–33). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0470870168.ch2

Donnelly, S., Brooks, P. J., & Homer, B. D. (2019). Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychonomic Bulletin & Review*, *26*, 1122–1147. https://doi.org/10.3758/s13423-019-01567-z

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300295

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Elliott, K. C. (2022). A taxonomy of transparency in science. *Canadian Journal of Philosophy*, *52*(3), 342–355. https://doi.org/10.1017/can.2020.21

Elston, D. M. (2021). Cherry picking, HARKing, and p-hacking. *Journal of the American Academy of Dermatology*. https://doi.org/https://doi.org/10.1016/j.jaad.2021.06.844

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fairfield, C. J., Harrison, E. M., & Wigmore, S. J. (2017). Duplicate publication bias weakens the validity of meta-analysis of immunosuppression after transplantation. *World Journal of Gastroenterology*, *23*(39), 7198. https://doi.org/10.3748/wjg.v23.i39.7198

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://doi.org/10.1177/2515245920952393

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA*, *105*(2), 203. https://doi.org/10.5195/jmla.2017.88

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *Observer*. https://www.psychologicalscience.org/observer/measurement-matters

Gall, T., Ioannidis, J. P., & Maniadis, Z. (2017). The credibility crisis in research: Can economics tools help? *PLoS Biology*, *15*(4), e2001846. https://doi.org/10.1371/journal.pbio.2001846

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*. Department of Statistics, Columbia University. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gibbons, M. (1999). Science's new social contract with society. *Nature*, *402*(6761), C81–C84. https://doi.org/10.1038/35011576

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., & Finos, L. (2022). *Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2210.02794

Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M.

(2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the netherlands. *PloS One*, *17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Haaf, J. M. (2020). *Conventional publication bias correction methods* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gv4tw

Haaf, J. M., Hoogeveen, S., Berkhout, S. W., Gronau, Q. F., & Wagenmakers, E.-J. (2020). *A bayesian multiverse analysis of many labs 4: Quantifying the evidence against mortality salience* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/cb9er

Hanel, P. H., & Zarzeczna, N. (2022). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion, Brain & Behavior*, 1–5. https://doi.org/10.1080/2153599X.2022.2070259

Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177. https://doi.org/10.1177/1745691620917678

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with r: A hands-on guide.* Chapman; Hall/CRC.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis= trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science*, *29*(3), 270–288. https://doi.org/10.

1177/0963662520902383

Heyman, T., & Vanpaemel, W. (2022). Multiverse analyses in the classroom. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2718

Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., et al. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, 1–47. https://doi.org/10.1080/2153599X.2022.2070255

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? *Psychological Methods*, *11*(2), 193. https://psycnet.apa.org/doi/10.1037/1082-989X.11.2.193

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P. (2017). Statistical biases in science communication: What we know about them and how they can be addressed. *The Oxford Handbook of the Science of Science Communication*, 102–110. https://doi.org/10.1093/oxfordhb/9780190497620.013.11

Judd, van L., N. (2023). *Ggrain: A rainclouds geom for 'ggplot2' (version 0.0.3)*. https://cran.r-project.org/package=ggrain

Kanters, S. (2022). Fixed- and random-effects models. In E. Evangelou & A. A. Veroniki (Eds.), *Meta-research: Methods and protocols* (pp. 41–65). Springer US.

https://doi.org/10.1007/978-1-0716-1566-9_3

Kathawalla, U.-K., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.18684

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *2*, 137–152. https://doi.org/10.1037/a0028086

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203/_4

Kim, J. H., & Choi, I. (2021). Choosing the level of significance: A decision-theoretic approach. *Abacus*, *57*(1), 27–71. https://doi.org/https://doi.org/10.1111/abac.12172

Krypotos, A.-M., Klein, R. A., & Jong, J. (2022). Resolving religious debates through a multiverse approach. *Religion, Brain & Behavior*, 1–3. https://doi.org/10.1080/2153599X.2022.2070261

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, 1–10. https://doi.org/10.1186/s40359-016-0126-3

Lakens, D., Page-Gould, E., Assen, M., Spellman, B., Schönbrodt, F., Hasselman, F., Corker, K., Grange, J., Sharples, A., Cavender, C., Augusteijn, H., Gerger, H., Locher, C., Miller, I., Anwari, F., & Scheel, A. (2017). *Examining the reproducibility of meta-analyses in psychology: A preliminary report*. https://doi.org/10.31222/osf.io/xfbjf

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/251524591877096

Lewandowsky, S., & Bishop, D. (2016). Research integrity: Don't let transparency damage science. *Nature*, *529*(7587), 459–461. https://doi.org/10.1038/529459a

Lishner, D. A. (2021). HARKing: Conceptualizations, harms, and two fundamental remedies. *Journal of Theoretical and Philosophical Psychology*, *41*(4), 248. https://doi.org/10.1037/teo0000182

Liu, Y. (2022). *Supporting reliable data analysis by evaluating all reasonable analytic decisions* [PhD thesis]. University of Washington.

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2021). Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19. https://psycnet.apa.org/doi/10.1037/1082-989X.7.1.19

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 1–14. https://doi.org/10.1177/25152459221080396

Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, *4*(3), 445–455. https://doi.org/10.1080/19439342.2012.711342

Martin, G., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, 523. https://doi.org/10.3389/fpsyg.2017.00523

Masur, P. K., & Scharkow, M. (2020). *Specr: Conducting and visualizing specification curve analyses (version 1.0.0)*. https://CRAN.R-project.org/package=specr

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Mayo-Wilson, E., Li, T., Fusco, N., Bertizzolo, L., Canner, J. K., Cowley, T., Doshi, P., Ehmsen, J., Gresham, G., Guo, N., et al. (2017). Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of Clinical Epidemiology*, *91*, 95–110. https://doi.org/10.1016/j.jclinepi.2017.07.014

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*(4), 386–401. https://psycnet.apa.org/doi/10.1037/1082-989X.11.4.386

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

Modecki, K. L., Low-Choy, S., Uink, B. N., Vernon, L., Correia, H., & Andrews, K.

(2020). Tuning into the real effect of smartphone use on parenting: A multiverse analysis. *Journal of Child Psychology and Psychiatry*, *61*(8), 855–865. https://doi.org/10.1111/jcpp.13282

Moring, B. (2017). *Research methods in psychology: Evaluating a world of information* (3rd ed.). WW Norton & Company.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Morse, J. M. (2010). "Cherry picking": Writing from thin data. *Qualitative Health Research*, *20*(1), 3–3. https://doi.org/10.1177/1049732309354285

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, *34*, 1–17. https://doi.org/10.1007/s10869-017-9524-7

National Academies of Sciences, Engineering, Medicine, et al. (2019). *Reproducibility and replicability in science*. National Academies Press.

Nuijten, M. B. (2019). Practical tools and strategies for researchers to increase replicability. *Developmental Medicine & Child Neurology*, *61*(5), 535–539. https://doi.org/10.1111/dmcn.14054

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

Patil, P., Peng, R. D., & Leek, J. T. (2016). A statistical definition for reproducibility and replicability. *BioRxiv*. https://doi.org/10.1101/066803

Pipal, C., Song, H., & Boomgaarden, H. G. (2022). If you have choices, why not choose (and share) all of them? A multiverse approach to understanding news engagement on social media. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2022.2036623

Plessen, C. Y., Karyotaki, E., & Cuijpers, P. (2022). Exploring the efficacy of psychological treatments for depression: A multiverse meta-analysis protocol. *BMJ Open*, *12*(1), e050197. https://doi.org/10.1136/bmjopen-2021-050197

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

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

Raj, A. T., Patil, S., Sarode, S., & Sarode, G. (2017). P-hacking. *The Journal of Contemporary Dental Practice*, *18*(8), 633–634. https://doi.org/10.5005/jp-journals-10024-2097

Rauvola, R. S., & Rudolph, C. W. (2023). Worker aging, control, and well-being: A specification curve analysis. *Acta Psychologica*, *233*, 103833. https://doi.org/10.1016/j.actpsy.2023.103833

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, *36*(1), 203–215. https://doi.org/10.1002/acp.3911

Rijnhart, J. J., Twisk, J. W., Deeg, D. J., & Heymans, M. W. (2021). Assessing the robustness of mediation analysis results using multiverse analysis. *Prevention Science*, *23*, 821–831. https://doi.org/10.1007/s11121-021-01280-1

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on i(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*, 1–9. https://doi.org/10.1186/1471-2288-8-79

Ruiter, J. de. (2019). Redefine or justify? Comments on the alpha debate. *Psychonomic Bulletin & Review*, *26*(2), 430–433. https://doi.org/10.3758/s13423-018-

1523-9

Rutjens, B. T., Heine, S. J., Sutton, R. M., & Van Harreveld, F. (2018). Attitudes towards science. In *Advances in experimental social psychology* (Vol. 57, pp. 125–165). Elsevier. https://doi.org/10.1016/bs.aesp.2017.08.001

Sarma, A., Kale, A., Moon, M., Taback, N., Chevalier, F., Hullman, J., & Kay, M. (2021). *Multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.1)*. https://github.com/MUCollective/multiverse

Scheufele, D. A. (2013). Communicating science in social settings. *Proceedings of the National Academy of Sciences*, *110*(supplement_3), 14040–14047. https://doi.org/10.1073/pnas.1213275110

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2019.1645

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Slez, A. (2019). The difference between instability and uncertainty: Comment on young and holsteen (2017). *Sociological Methods & Research*, *48*(2), 400–430. https:

//doi.org/10.1177/0049124117729704

Spineli, L. M., & Pandis, N. (2020). The importance of careful selection between fixed-effect and random-effects models. *American Journal of Orthodontics and Dentofacial Orthopedics*, *157*(3), 432–433. https://doi.org/10.1016/j.ajodo.2019.12.003

Srivastava, S. (2018). *Sound inference in complicated research: A multi-strategy approach* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/bwr48

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In *Publication bias in meta-analysis* (pp. 73–98). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0470870168.ch5

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, *8*, 862. https://doi.org/10.3389/fpsyg.2017.00862

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303. https://doi.org/10.1177/096228020101000404

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., Rooij, I. van, Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Ter Schure, J., & Grünwald, P. D. (2019). Accumulation bias in meta-analysis: The need to consider time in error control. In *arXiv* [Preprint]. https://doi.org/10.48550/arXiv.1905.13494

Van Rooij, I. (2019). *Psychological science needs theory development before preregistration.* https://featuredcontent.psychonomic.org/psychological-science-needs-theory-development-before-preregistration/

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. https://doi.org/https://doi.org/10.1002/sim.2514

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package (version 4.0.0). *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? *Zeitschrift für Psychologie*, *227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Wessel, I., Albers, C. J., Zandstra, A. R. E., & Heininga, V. E. (2020). A multiverse analysis of early attempts to replicate memory suppression with the think/no-think task. *Memory*, *28*(7), 870–887. https://doi.org/10.1080/09658211.2020.1797095

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832. https://doi.org/10.3389/fpsyg.2016.0183

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Winer, E. S., Jordan, D. G., & Collins, A. C. (2019). Conceptualizing anhedonias and implications for depression treatments. *Psychology Research and Behavior Management*, *12*, 325. https://doi.org/10.2147%2FPRBM.S159260

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3–40. https://doi.org/10.1177/0049124115610347

Appendix

The multiverse matrix

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.60	0.40	fit fixed	d_{ppc2}	0.24	0.02	13.06	<0.001	0.20	0.27
0.60	0.40	fit fixed	τ^2	0.00					
0.60	0.40	fit random	d_{ppc2}	0.42	0.06	7.48	<0.001	0.31	0.53
0.60	0.40	fit random	τ^2	0.19	0.04				
0.60	0.45	fit fixed	d_{ppc2}	0.24	0.02	12.99	<0.001	0.20	0.27
0.60	0.45	fit fixed	τ^2	0.00					
0.60	0.45	fit random	d_{ppc2}	0.42	0.06	7.50	<0.001	0.31	0.53
0.60	0.45	fit random	τ^2	0.19	0.04				
0.60	0.50	fit fixed	d_{ppc2}	0.24	0.02	12.92	<0.001	0.20	0.27
0.60	0.50	fit fixed	τ^2	0.00					
0.60	0.50	fit random	d_{ppc2}	0.42	0.06	7.51	<0.001	0.31	0.53
0.60	0.50	fit random	τ^2	0.18	0.04				
0.60	0.55	fit fixed	d_{ppc2}	0.24	0.02	12.85	<0.001	0.20	0.27
0.60	0.55	fit fixed	τ^2	0.00					
0.60	0.55	fit random	d_{ppc2}	0.42	0.06	7.52	<0.001	0.31	0.52
0.60	0.55	fit random	τ^2	0.18	0.04				
0.60	0.60	fit fixed	d_{ppc2}	0.24	0.02	12.79	<0.001	0.20	0.27
0.60	0.60	fit fixed	τ^2	0.00					
0.60	0.60	fit random	d_{ppc2}	0.42	0.06	7.54	<0.001	0.31	0.52
0.60	0.60	fit random	τ^2	0.18	0.04				
0.60	0.65	fit fixed	d_{ppc2}	0.23	0.02	12.73	<0.001	0.20	0.27

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.60	0.65	fit fixed	τ^2	0.00					
0.60	0.65	fit random	d_{ppc2}	0.41	0.05	7.56	<0.001	0.31	0.52
0.60	0.65	fit random	τ^2	0.18	0.04				
0.60	0.70	fit fixed	d_{ppc2}	0.23	0.02	12.67	<0.001	0.20	0.27
0.60	0.70	fit fixed	τ^2	0.00					
0.60	0.70	fit random	d_{ppc2}	0.41	0.05	7.58	<0.001	0.31	0.52
0.60	0.70	fit random	τ^2	0.18	0.04				
0.60	0.75	fit fixed	d_{ppc2}	0.23	0.02	12.62	<0.001	0.20	0.27
0.60	0.75	fit fixed	τ^2	0.00					
0.60	0.75	fit random	d_{ppc2}	0.41	0.05	7.61	<0.001	0.31	0.52
0.60	0.75	fit random	τ^2	0.17	0.04				
0.60	0.80	fit fixed	d_{ppc2}	0.23	0.02	12.56	<0.001	0.20	0.27
0.60	0.80	fit fixed	τ^2	0.00					
0.60	0.80	fit random	d_{ppc2}	0.41	0.05	7.65	<0.001	0.31	0.52
0.60	0.80	fit random	τ^2	0.17	0.04				
0.65	0.40	fit fixed	d_{ppc2}	0.24	0.02	13.76	<0.001	0.20	0.27
0.65	0.40	fit fixed	τ^2	0.00					
0.65	0.40	fit random	d_{ppc2}	0.42	0.06	7.52	<0.001	0.31	0.53
0.65	0.40	fit random	τ^2	0.19	0.04				
0.65	0.45	fit fixed	d_{ppc2}	0.23	0.02	13.68	<0.001	0.20	0.27
0.65	0.45	fit fixed	τ^2	0.00					
0.65	0.45	fit random	d_{ppc2}	0.42	0.06	7.53	<0.001	0.31	0.53
0.65	0.45	fit random	τ^2	0.19	0.04				
0.65	0.50	fit fixed	d_{ppc2}	0.23	0.02	13.61	<0.001	0.20	0.27
0.65	0.50	fit fixed	τ^2	0.00					
0.65	0.50	fit random	d_{ppc2}	0.42	0.06	7.54	<0.001	0.31	0.53
0.65	0.50	fit random	τ^2	0.19	0.04				
0.65	0.55	fit fixed	d_{ppc2}	0.23	0.02	13.54	<0.001	0.20	0.27
0.65	0.55	fit fixed	τ^2	0.00					
0.65	0.55	fit random	d_{ppc2}	0.42	0.06	7.56	<0.001	0.31	0.52
0.65	0.55	fit random	τ^2	0.19	0.04				

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.65	0.60	fit fixed	d_{ppc2}	0.23	0.02	13.48	<0.001	0.20	0.27
0.65	0.60	fit fixed	τ^2	0.00					
0.65	0.60	fit random	d_{ppc2}	0.42	0.05	7.57	<0.001	0.31	0.52
0.65	0.60	fit random	τ^2	0.18	0.04				
0.65	0.65	fit fixed	d_{ppc2}	0.23	0.02	13.42	<0.001	0.20	0.27
0.65	0.65	fit fixed	τ^2	0.00					
0.65	0.65	fit random	d_{ppc2}	0.42	0.05	7.59	<0.001	0.31	0.52
0.65	0.65	fit random	τ^2	0.18	0.04				
0.65	0.70	fit fixed	d_{ppc2}	0.23	0.02	13.36	<0.001	0.20	0.26
0.65	0.70	fit fixed	τ^2	0.00					
0.65	0.70	fit random	d_{ppc2}	0.41	0.05	7.61	<0.001	0.31	0.52
0.65	0.70	fit random	τ^2	0.18	0.04				
0.65	0.75	fit fixed	d_{ppc2}	0.23	0.02	13.30	<0.001	0.20	0.26
0.65	0.75	fit fixed	τ^2	0.00					
0.65	0.75	fit random	d_{ppc2}	0.41	0.05	7.64	<0.001	0.31	0.52
0.65	0.75	fit random	τ^2	0.18	0.04				
0.65	0.80	fit fixed	d_{ppc2}	0.23	0.02	13.24	<0.001	0.20	0.26
0.65	0.80	fit fixed	τ^2	0.00					
0.65	0.80	fit random	d_{ppc2}	0.41	0.05	7.68	<0.001	0.31	0.52
0.65	0.80	fit random	τ^2	0.17	0.04				
0.70	0.40	fit fixed	d_{ppc2}	0.23	0.02	14.60	<0.001	0.20	0.26
0.70	0.40	fit fixed	τ^2	0.00					
0.70	0.40	fit random	d_{ppc2}	0.42	0.06	7.55	<0.001	0.31	0.53
0.70	0.40	fit random	τ^2	0.20	0.04				
0.70	0.45	fit fixed	d_{ppc2}	0.23	0.02	14.52	<0.001	0.20	0.26
0.70	0.45	fit fixed	τ^2	0.00					
0.70	0.45	fit random	d_{ppc2}	0.42	0.06	7.56	<0.001	0.31	0.53
0.70	0.45	fit random	τ^2	0.19	0.04				
0.70	0.50	fit fixed	d_{ppc2}	0.23	0.02	14.45	<0.001	0.20	0.26
0.70	0.50	fit fixed	τ^2	0.00					
0.70	0.50	fit random	d_{ppc2}	0.42	0.06	7.57	<0.001	0.31	0.53

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.70	0.50	fit random	τ^2	0.19	0.04				
0.70	0.55	fit fixed	d_{ppc2}	0.23	0.02	14.38	<0.001	0.20	0.26
0.70	0.55	fit fixed	τ^2	0.00					
0.70	0.55	fit random	d_{ppc2}	0.42	0.06	7.59	<0.001	0.31	0.53
0.70	0.55	fit random	τ^2	0.19	0.04				
0.70	0.60	fit fixed	d_{ppc2}	0.23	0.02	14.31	<0.001	0.20	0.26
0.70	0.60	fit fixed	τ^2	0.00					
0.70	0.60	fit random	d_{ppc2}	0.42	0.05	7.61	<0.001	0.31	0.52
0.70	0.60	fit random	τ^2	0.19	0.04				
0.70	0.65	fit fixed	d_{ppc2}	0.23	0.02	14.25	<0.001	0.20	0.26
0.70	0.65	fit fixed	τ^2	0.00					
0.70	0.65	fit random	d_{ppc2}	0.42	0.05	7.62	<0.001	0.31	0.52
0.70	0.65	fit random	τ^2	0.19	0.04				
0.70	0.70	fit fixed	d_{ppc2}	0.23	0.02	14.18	<0.001	0.20	0.26
0.70	0.70	fit fixed	τ^2	0.00					
0.70	0.70	fit random	d_{ppc2}	0.42	0.05	7.65	<0.001	0.31	0.52
0.70	0.70	fit random	τ^2	0.18	0.04				
0.70	0.75	fit fixed	d_{ppc2}	0.23	0.02	14.12	<0.001	0.20	0.26
0.70	0.75	fit fixed	τ^2	0.00					
0.70	0.75	fit random	d_{ppc2}	0.41	0.05	7.68	<0.001	0.31	0.52
0.70	0.75	fit random	τ^2	0.18	0.04				
0.70	0.80	fit fixed	d_{ppc2}	0.23	0.02	14.06	<0.001	0.19	0.26
0.70	0.80	fit fixed	τ^2	0.00					
0.70	0.80	fit random	d_{ppc2}	0.41	0.05	7.72	<0.001	0.31	0.52
0.70	0.80	fit random	τ^2	0.18	0.04				
0.75	0.40	fit fixed	d_{ppc2}	0.23	0.01	15.63	<0.001	0.20	0.26
0.75	0.40	fit fixed	τ^2	0.00					
0.75	0.40	fit random	d_{ppc2}	0.42	0.06	7.58	<0.001	0.31	0.53
0.75	0.40	fit random	τ^2	0.20	0.04				
0.75	0.45	fit fixed	d_{ppc2}	0.23	0.01	15.55	<0.001	0.20	0.25
0.75	0.45	fit fixed	τ^2	0.00					

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.75	0.45	fit random	d_{ppc2}	0.42	0.06	7.59	<0.001	0.31	0.53
0.75	0.45	fit random	τ^2	0.20	0.04				
0.75	0.50	fit fixed	d_{ppc2}	0.23	0.01	15.48	<0.001	0.20	0.25
0.75	0.50	fit fixed	τ^2	0.00					
0.75	0.50	fit random	d_{ppc2}	0.42	0.06	7.60	<0.001	0.31	0.53
0.75	0.50	fit random	τ^2	0.20	0.04				
0.75	0.55	fit fixed	d_{ppc2}	0.22	0.01	15.41	<0.001	0.20	0.25
0.75	0.55	fit fixed	τ^2	0.00					
0.75	0.55	fit random	d_{ppc2}	0.42	0.05	7.62	<0.001	0.31	0.53
0.75	0.55	fit random	τ^2	0.20	0.04				
0.75	0.60	fit fixed	d_{ppc2}	0.22	0.01	15.34	<0.001	0.20	0.25
0.75	0.60	fit fixed	τ^2	0.00					
0.75	0.60	fit random	d_{ppc2}	0.42	0.05	7.64	<0.001	0.31	0.52
0.75	0.60	fit random	τ^2	0.19	0.04				
0.75	0.65	fit fixed	d_{ppc2}	0.22	0.01	15.27	<0.001	0.19	0.25
0.75	0.65	fit fixed	τ^2	0.00					
0.75	0.65	fit random	d_{ppc2}	0.42	0.05	7.66	<0.001	0.31	0.52
0.75	0.65	fit random	τ^2	0.19	0.04				
0.75	0.70	fit fixed	d_{ppc2}	0.22	0.01	15.21	<0.001	0.19	0.25
0.75	0.70	fit fixed	τ^2	0.00					
0.75	0.70	fit random	d_{ppc2}	0.42	0.05	7.68	<0.001	0.31	0.52
0.75	0.70	fit random	τ^2	0.19	0.04				
0.75	0.75	fit fixed	d_{ppc2}	0.22	0.01	15.15	<0.001	0.19	0.25
0.75	0.75	fit fixed	τ^2	0.00					
0.75	0.75	fit random	d_{ppc2}	0.41	0.05	7.71	<0.001	0.31	0.52
0.75	0.75	fit random	τ^2	0.19	0.04				
0.75	0.80	fit fixed	d_{ppc2}	0.22	0.01	15.08	<0.001	0.19	0.25
0.75	0.80	fit fixed	τ^2	0.00					
0.75	0.80	fit random	d_{ppc2}	0.41	0.05	7.75	<0.001	0.31	0.52
0.75	0.80	fit random	τ^2	0.18	0.04				
0.80	0.40	fit fixed	d_{ppc2}	0.22	0.01	16.97	<0.001	0.20	0.25

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.80	0.40	fit fixed	τ^2	0.00					
0.80	0.40	fit random	d_{ppc2}	0.42	0.06	7.61	<0.001	0.31	0.53
0.80	0.40	fit random	τ^2	0.21	0.04				
0.80	0.45	fit fixed	d_{ppc2}	0.22	0.01	16.89	<0.001	0.19	0.25
0.80	0.45	fit fixed	τ^2	0.00					
0.80	0.45	fit random	d_{ppc2}	0.42	0.06	7.62	<0.001	0.31	0.53
0.80	0.45	fit random	τ^2	0.20	0.04				
0.80	0.50	fit fixed	d_{ppc2}	0.22	0.01	16.81	<0.001	0.19	0.25
0.80	0.50	fit fixed	τ^2	0.00					
0.80	0.50	fit random	d_{ppc2}	0.42	0.05	7.63	<0.001	0.31	0.53
0.80	0.50	fit random	τ^2	0.20	0.04				
0.80	0.55	fit fixed	d_{ppc2}	0.22	0.01	16.73	<0.001	0.19	0.24
0.80	0.55	fit fixed	τ^2	0.00					
0.80	0.55	fit random	d_{ppc2}	0.42	0.05	7.65	<0.001	0.31	0.53
0.80	0.55	fit random	τ^2	0.20	0.04				
0.80	0.60	fit fixed	d_{ppc2}	0.22	0.01	16.66	<0.001	0.19	0.24
0.80	0.60	fit fixed	τ^2	0.00					
0.80	0.60	fit random	d_{ppc2}	0.42	0.05	7.67	<0.001	0.31	0.53
0.80	0.60	fit random	τ^2	0.20	0.04				
0.80	0.65	fit fixed	d_{ppc2}	0.22	0.01	16.59	<0.001	0.19	0.24
0.80	0.65	fit fixed	τ^2	0.00					
0.80	0.65	fit random	d_{ppc2}	0.42	0.05	7.68	<0.001	0.31	0.52
0.80	0.65	fit random	τ^2	0.20	0.04				
0.80	0.70	fit fixed	d_{ppc2}	0.22	0.01	16.53	<0.001	0.19	0.24
0.80	0.70	fit fixed	τ^2	0.00					
0.80	0.70	fit random	d_{ppc2}	0.42	0.05	7.71	<0.001	0.31	0.52
0.80	0.70	fit random	τ^2	0.19	0.04				
0.80	0.75	fit fixed	d_{ppc2}	0.22	0.01	16.46	<0.001	0.19	0.24
0.80	0.75	fit fixed	τ^2	0.00					
0.80	0.75	fit random	d_{ppc2}	0.42	0.05	7.74	<0.001	0.31	0.52
0.80	0.75	fit random	τ^2	0.19	0.04				

The multiverse matrix *(continued)*

rmorris	ragg	model	term	estimate	std.error	statistic	p.value	conf.low	conf.high
0.80	0.80	fit fixed	d_{ppc2}	0.22	0.01	16.39	<0.001	0.19	0.24
0.80	0.80	fit fixed	τ^2	0.00					
0.80	0.80	fit random	d_{ppc2}	0.41	0.05	7.78	<0.001	0.31	0.52
0.80	0.80	fit random	τ^2	0.19	0.04				