UNIVERSITY OF PADOVA

Department of General Psychology

Master Degree in
Neuroscience and Neuropsychological Rehabilitation

Final Dissertation

Improving the Interpretation of Effect Sizes through Modeling and Simulation

Supervisor:
Professor Gianmarco Altoè

Co-supervisor:
Professor Daniel Lakens

Candidate: Laura Sità
Student ID number: 2085633

Academic year 2024/2025

Table of Contents

Summary

Psychology is currently facing a substantial credibility crisis and restoring trust in the field necessitates the adoption of rigorous research practices and transparency (Ioannidis, 2005). This issue relates to the formalization of hypotheses and associated preplanned statistical analysis and the interpretation of findings resulting respectively in the theory crisis and the replication findings. It is advisable for researchers to address issues related to theory building and replicability, as these problems undermine the credibility of scientific evidence. One approach to addressing these issues is through the introduction and widespread adoption of more appropriate research practices to restore the trust and reliability essential for fields like psychology, where the ultimate goal is to improve people's lives. Accordingly, we explored potential remedies to these challenges, emphasizing research practices that have recently been identified as crucial for advancement in psychological research: the development of appropriate statistical models based on theoretical considerations (Fried, 2020) and the pursuit of more theoretically meaningful interpretations of effect sizes (Anvari & Lakens, 2021).

This dissertation begins by outlining the crises currently impacting psychological research, examining their root causes and suggesting how developing sound statistical models and simulations prior to empirical testing - along with more rigorous interpretations of key research parameters, such as the effect size - may offer effective remedies. Chapter Two will focus on the concept of effect size, providing a formal definition, describing various methods for its calculation and highlighting the limitations of one

of its most widely used indices, Cohen's *d*. Chapter Three introduces the specific scenario in which we aim to apply the proposed framework, which we hope will encourage researchers to adopt more robust research practices. This framework is designed to guide researchers in conceptualizing and modeling effect sizes, as well as conducting simulations prior to data collection. The framework will be implemented within the simulations of different scenarios in Chapter Four, followed by application to a real case study in Chapter Five.

This thesis was written in R markdown.

Chapter 1

Crises in psychology

> *"If statisticians agree on one thing, it is that scientific inference should not be made mechanically."*
>
> — Gigerenzer and Marewski, 2015, p. 422

In this dissertation, we begin by outlining the crises currently affecting psychological research. After providing definitions and a discussion of their underlying causes, we illustrate how the theory and replication crises may be addressed by improving particular research practices. Specifically, we will explain how the use of models and simulations prior to testing, along with more meaningful interpretations of key parameters in psychological research - such as the effect size - represent the improvements we aim to address with the framework proposed in this thesis.

1.1 Crises within confirmatory research

Psychology is a challenging discipline. Empirical data are noisy, formal theory is scarce, and the processes of interest (e.g., attention, jealousy, loss aversion) cannot be observed directly. Nevertheless, psychologists have managed to generate many key insights about human cognition and behavior (Wagenmakers et al., 2012). Nevertheless, several years

ago, Ioannidis (2005) famously argued that "most published research findings are false" (p. 696) and psychology is not immune to this accusation. The consequences of crises that have been documented for more than half a century have now become so relevant that they can no longer be overlooked. Given that these crises pertain to scientific modes of investigation, it has become imperative to address the issue with urgency.

The scientific practices impacted by these crises include both exploratory and confirmatory research methodologies. By way of definition, exploratory research is about free and flexible examination of a dataset without predetermined notions, with the goal of detecting patterns and associations that may eventually lead to the generation of new hypotheses. In contrast, confirmatory research attempts to evaluate a clear hypothesis that may or may not be confirmed by the data collected (Kimmelman et al., 2014). In confirmatory studies, researchers must pre-plan and formalize hypotheses and consequently test them (Wagenmakers et al., 2012).

Although scientists often think of experiments in the context of confirmation, philosophers of science have emphasized the role of exploratory experiments in theory development. In exploratory experiments, researchers vary a large number of parameters without a priori predictions of their effects (although some prior knowledge of plausible parameters is necessary), look for stable empirical patterns and infer rules from these patterns (Scheel et al., 2021). For example, exploratory experimentation is widely used in psychophysics to establish law-like relationships. By focusing primarily on confirmatory research and jumping straight to the hypothesis test, psychologists too often neglect the groundwork that is necessary to ensure a sound link between the test and the tested theory (Scheel et al., 2021). The most crucial aspect is to maintain transparency, clarity and rigor regarding the type of analysis conducted, thereby ensuring academic integrity. There is no issue with exploratory analysis, provided it is explicitly acknowledged as such (Wagenmakers et al., 2012). Both exploratory and confirmatory research hold substantial value in scientific inquiry and are most effective when used

in conjunction, with exploratory research informing subsequent confirmatory studies. Accordingly, we state that a confirmatory study has been chosen for this work, which will be presented in the fifth chapter.

The three main aspects of confirmatory research are

1. formalization of hypotheses and associated pre-planned statistical analysis;
2. operationalization of the variables of interest;
3. interpretation of (statistical) findings.

Each of these aspects is accompanied by evidence of a crisis. These crises are not novel, as numerous articles - dating back to the 1950s (e.g., Meehl, 1967) - have been published on the subject, even though with limited impact on the scientific community. In the following sections of this paragraph, we will provide a more in-depth analysis of these crises and their underlying causes, with particular emphasis on the crisis associated with the first aspect, namely the theory crisis, and the crisis related to the third aspect, commonly referred to as the replication crisis.

1.1.1 Theory crisis: definition and causes

Meehl argued in 1978 that theories in psychology come and go, with little cumulative progress (Meehl, 1978). This assessment still holds, as also evidenced by increasingly common claims that psychology is facing a "theory crisis" (Eronen & Bringmann, 2021). According to this claim, psychological theories are in general of poor quality and the focus in psychology should shift more toward developing better theories instead of (just) improving statistical techniques and practices and performing more replication studies. Stated otherwise, psychologists should invest more in theory building.

The main factors that cause this crisis are the widespread amount of vague and not testable theories, the lack of mathematical formalization of hypotheses and the mismatch between theoretical and statistical models: it is common to observe inade-

quate formalization of theoretical models and frequently statistical models that fail to adequately test researchers' hypotheses. All these malpractices can be attributed to the common issue of researchers failing to fully commit to a pre-specified analysis plan prior to examining the data. This approach increases the likelihood that the reported findings are fictional and, consequently, non-replicable (Wagenmakers et al., 2012).

1.1.2 Replication crisis: definition and causes

It is universally acknowledged that reproducibility and replicability are fundamental characteristics of scientific studies. However, many scientists often use these terms interchangeably, yet, it is useful and essential to understand the distinction. Patil et al. (2016) provide informal definitions for these key scientific terms. 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. Replicability crisis started examining psychological papers and noticing that most current published research findings are false, because sometimes refuted by subsequent evidence, with ensuing confusion and disappointment (Ioannidis, 2005). Serious worries have been voiced concerning a "replicability crisis" in many biomedical as well as social sciences; this crisis of confidence is fueled by the observation that numerous established findings may correspond to false positives that cannot be reproduced (Gall et al., 2017).

The replicability crisis has uncovered a number of problematic issues, including the adoption of questionable research practices. According to Banks et al. (2016), questionable research practices are 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). These practices are largely responsible for provoking a broader distrust in scientific evidence, contributing to what is commonly referred to as the credibility crisis. Credibility crisis indicates the lack of confidence in results that characterizes research, especially in the fields of

social and biomedical sciences (Gall et al., 2017). This distrust is fueled by substantial evidence showing that many research studies fail to produce replicable results. Replicability is crucial for building confidence and credibility in science. However, over the past two decades, the field of psychological science has faced an unprecedented replicability crisis, which has significantly undermined the foundations of trust in scientific findings. (Aarts et al., 2015).

Some of the potential reasons for mistrust and credibility crisis consist in flawed practices including bending the data to produce publishable outcomes, misusing statistical tools and failing to bring justification when necessary, and using invalid measurements. One of the key aspects of the crisis that this thesis will address pertains to the calculation, use and interpretation of the effect size, given its relevance in the social sciences, particularly within psychological research. Effect size is a quantitative measure that describes the strength of a phenomenon and reflects its practical or clinical significance (Lakens, 2022). Given its importance, it is crucial to address and improve the malpractices associated with it. Calculating and reporting an observed effect size is merely the first step in the process; while necessary, it is not sufficient on its own. Some studies report effect sizes that are either too small - an issue that partly arises from the misguided belief that psychological phenomena are driven by many small, additive effects (Primbs et al., 2023) - or too large to be meaningful. Consequently, these effect sizes often turn out to be too small to replicate or too large to plausibly be attributed to the proposed theoretical mechanisms (Hilgard, 2021). On the contrary, it is often desirable to be able to conclude that the effect size in one study is larger than a minimally important difference and matters in practice (Otgaar et al., 2022). This issue arises because researchers often consider effect sizes only at the end of their study, leading to superficial or meaningless interpretations after analyzing the data. In short, these poor interpretations contribute to reporting effect sizes that are not replicable, which substantially exacerbates the credibility crisis.

In 1962, Cohen highlighted a persistent issue in psychological research that remains relevant today (Cohen, 1962). Researchers appeared to overlook the statistical power of their studies, which refers to the probability of correctly rejecting the null hypothesis given the presence of a true experimental effect in the population. As a result, underpowered studies became increasingly prevalent, thereby contaminating the published literature in psychology. Underpowered studies not only contribute to a confusing body of literature but also result in biased estimates of effect sizes. Moreover, researchers may have felt little incentive to enhance the power of their studies, as the practice of testing multiple hypotheses - now a common approach - often provided a reasonable likelihood of obtaining at least one statistically significant result (Maxwell, 2004).

Overall, a lack of thorough understanding of these issues will lead to the continued repetition of these flawed practices, ultimately undermining the reliability of scientific conclusions.

1.2 Possible remedies

In the previous section, 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 pertinent question is whether it is inevitable that most research findings are erroneous or if there are ways to improve the situation. A considerable challenge is the inherent impossibility of achieving absolute certainty regarding the truth of any research question. In this context, the ideal "gold standard" remains unattainable. Nevertheless, several strategies exist to enhance the reliability of research findings (Ioannidis, 2005).

1.2.1 Thinking (and modeling) before testing

A disconcerting reality that undermines the foundation of academic psychology is that, with few exceptions, psychologists often do not commit to a method of data analysis prior to examining the actual data. This lack of pre-commitment creates an inclination to adjust the analysis in response to the data in order to achieve a desired result. Such practices compromise the validity of the interpretation of commonly used statistical tests (Wagenmakers et al., 2012). As Cummins stated: "In psychology, we are overwhelmed with things to explain, and somewhat underwhelmed by things to explain them with" (Cummins, 2000, p. 120). Sometimes, research suffers from poor theoretical model formalization and the inappropriate use of statistical models in hypothesis testing. The relevance of these mechanisms can vary depending on the context and the phenomena being studied (Anvari et al., 2023). Conversely, well-designed models may be statistically sound but fail to offer meaningful insights or explanations, making them practically useless. Therefore, aligning theoretical models with appropriate statistical methods is essential for effectively testing theories. A renewed emphasis on theoretical psychology and the formalization of theories can provide a path forward (Fried, 2020). To effectively connect theories to data, it is essential to use statistical models that align with the theoretical assumptions. Selecting an appropriate statistical model that matches the assumptions of the theory is crucial for accurately applying data to theoretical constructs. This alignment can be achieved only if experts in the relevant field and statisticians collaborate from the outset, prior to data collection. Such early collaboration offers a valid opportunity to improve the situation and address the theory crisis.

Considering methodological issues during the experimental design phase is insufficient. Researchers must also ensure transparency in their procedures to contribute to scientific progress within an open science framework, an approach widely recognized as

one of the most effective remedies for addressing the credibility crisis. Open science is an umbrella term referring to a variety of practices and principles to ensure transparency, credibility, reproducibility, and accessibility.

A potential tool that addresses certain aspects of transparency is preregistration, which refers to a dated document containing research questions, the hypotheses, the method and the analysis plan that is published before data collection (Ummul-Kiram et al., 2021). While preregistration can reduce the likelihood of false findings and is considered essential for a study to be classified as a confirmatory study (Wagenmakers et al., 2012), some researchers have found ways to circumvent its integrity. This is done by referencing the preregistration in their final reports only when the results are significant, while disregarding it altogether if the outcomes are not. A more robust alternative is the use of Registered Reports, which follow the process outlined in figure 1.1. Registered Reports are a publishing format that prioritizes the research question and methodological rigor by subjecting the study to peer review before data collection. High-quality protocols are provisionally accepted for publication, provided the authors adhere to the registered methodology. Since studies are accepted in advance, this practice effectively shifts incentive structures, promoting replication efforts and the reporting of results regardless of their statistical significance (Nosek & Lakens, 2016).

Figure 1.1: Procedure for conducting Registered Reports, emphasizing the inclusion of peer review prior to data collection and results. This illustration is taken from Center for Open Science, (n.d.), Registered reports, https://www.cos.io/initiatives/registered-reports

1.2.2 Meaningful interpretations of the effect size

Scientific progress in quantitative psychology relies on our ability to derive meaningful statistical inferences from empirical data. For over a century, researchers have emphasized the importance of not only interpreting the statistical significance of effects but also assessing their practical significance (Boring, 1919). Therefore, it is crucial to focus on effect size and its meaningful interpretation. As Pek & Flora (2018) highlight, "the meaningful interpretation of reported effect sizes is the essence of what contributes to our science" (p. 221), while, in practice, the meaningful interpretation of effect sizes is often lacking in psychology.

We will now examine three primary approaches to effect size that can serve as remedies for the replicability crisis. Firstly, determining a priori the effect size large enough to be practically meaningful (namely the Smallest Effect Size of Interest - SESOI) (Anvari & Lakens, 2021), as well as thinking about the expected effect size in general, can improve the credibility of research to a reasonable extent because it provides the possibility to design falsifiable studies (Anvari & Lakens, 2021), which is a fundamental characteristic of replicability. Secondly, effect sizes from previous studies can be used, together with required significance level and statistical power, when planning a new study in an a-priori power analysis (Lakens, 2022), that is to say the calculation used to estimate the smallest sample size needed for an experiment before collecting the data. Finally, since in very large samples p-values tend to approach zero rapidly, relying solely on p-values can lead researchers to erroneously claim support for results that lack practical significance (Lin et al., 2013), potentially due to biased research practices. Therefore, it is increasingly important for researchers to assess whether observed effects are theoretically or practically significant. This practice helps to prevent the common misinterpretation of 'statistical significance' as 'practical significance,' given the rise of big data and the uptake of large-scale collaborative projects (Anvari & Lakens, 2021).

A key aspect shared by these remedies consists in the emphasis on addressing effect sizes from the outset of the study, rather than only considering them during the analysis of collected data. Incorporating effect sizes into the study design enables researchers to formulate predictions that, particularly when simulated, can lead to more meaningful interpretations of the results derived from data analysis.

Furthermore, improvements in the field of effect sizes can be achieved by addressing the statistical power of studies. As previously noted, underpowered studies are prevalent and one key issue with such studies is that the probability of detecting an effect, should it exist, is low. More critically, if a statistically significant result is obtained in an underpowered study, the effect size associated with the observed p-value may be disproportionately large or "too big to be true" (Altoè et al., 2020). This highlights the close interrelationship between statistical power and effect size, underscoring the need for a framework that addresses both issues. For psychology to continue building a coherent and reliable body of scientific literature, it is essential to place greater emphasis on the role of power in study design and result interpretation.

1.3 Proposed framework

Given the pressing issues currently impacting research progress, particularly within psychology, a crucial next step is to consider effect size not merely as a numerical value but in the context of how it can be integrated into a theoretical model. Researchers should explicitly model the predicted effect of the independent variable on the dependent variable and test this model through their hypotheses. This approach would, for example, help prevent the identification of effect sizes that are implausibly large, as noted by Hilgard (2021). Furthermore, we can enhance our study by conducting a computer simulation prior to any data collection, as recommended by Gelman (2024). This approach not only reveals designs that may be too noisy to detect main effects or

interactions of interest but also sharpens our focus by requiring us to make informed assumptions about the structure and magnitude of effects. Simulations can ensure that our measurements effectively address the underlying construct of interest and that our experimental design considers realistic effect sizes (Gelman, 2024). An important feature to incorporate into simulations could be the provision of graphical representations of the simulated outcomes. This allows researchers to visually examine the effects under study, facilitating reflection on the results. Additionally, after completing the analysis of real data, researchers can compare the simulated results with the actual outcomes, aiding in the interpretation by providing a visual context for the phenomena rather than relying solely on numerical values.

Confirmatory research should be conducted according to the following steps:

1. based on theory, formulate and formalize hypotheses;
2. identify the relevant variables and how to measure them;
3. plan the research design, sample size and statistical analysis;
4. conduct a computer simulation and calculate the statistical power;
5. collect data;
6. test the hypotheses;
7. interpret and share results.

This process includes, as its initial step, the implementation of all aforementioned considerations. As Tukey (1987) aptly remarked, “Finding the question is often more important than finding the answer” (p. 511). This statement underscores the critical importance of the study design phase. It is therefore advisable to include expected effect sizes in this initial stage. Thoroughly considering and modeling effect sizes at this early stage together with preceding any data collection with a computer simulation is essential for ensuring that hypotheses are falsifiable and for adhering to best practices in research.

Integrating these two aspects would significantly enhance research in psychology. By proactively considering effect sizes and modeling them before testing their data, quantitative psychologists can design more informative and efficient studies. Moreover, simulations can assist researchers in critically reflecting on their intended data analysis before the actual data collection occurs and, by incorporating visual representations into the simulations, achieving more meaningful interpretations of the study's results. This whole approach will help counteract misleading claims of significance, enable researchers to statistically falsify predictions, improve the interpretation of replication studies and provide psychological scientists with a conceptual framework for interpreting effect sizes. As will be further detailed in the third chapter, the proposed framework aims to preemptively address fundamental design flaws that cannot be corrected after data collection, with the goal of ensuring that the data can inform the statistical hypothesis being tested, thereby addressing a critical need recognized by researchers (Lakens, 2023). After a century of highlighting these issues, it is imperative to address these challenges to advance both the field of psychology and the broader research community.

1.4 Aims of the current study

In this chapter, we examined the problems associated with confirmatory studies, which contribute to the proliferation of large quantities of low-quality research (Ioannidis, 2005). This issue relates to the formalization of hypotheses and associated pre-planned statistical analysis and the interpretation of findings resulting respectively in the theory crisis and the replication findings. It is advisable for researchers to address issues related to theory building and replicability, as these problems undermine the credibility of scientific evidence. One approach to addressing these issues is through the introduction and widespread adoption of more appropriate research practices to restore the trust

and reliability essential for fields like psychology, where the ultimate goal is to improve people's lives. Accordingly, we explored potential remedies to these challenges, emphasizing research practices that have recently been identified as crucial for advancement in psychological research: the development of appropriate statistical models based on theoretical considerations (Fried, 2020) and the pursuit of more theoretically meaningful interpretations of effect sizes (Anvari & Lakens, 2021).

In the present dissertation we aim to propose a framework, as a solution to these issues, designed to guide researchers in conceptualizing and modeling effect sizes, as well as conducting data simulations prior to data collection. This framework is presented as a structured sequence of steps that researchers can implement throughout the research process, with the goal of improving both the quality and reliability of their studies. Moreover, this framework is designed to assist psychologists in thoroughly considering various critical aspects of the study (e.g., sample size, variability of predicted effect sizes, statistical power) through visual representations, emphasizing the importance of moving beyond merely reporting numerical values in the results. After testing these hypotheses through data analysis, researchers can then reflect on the actual results obtained versus the expected outcomes. Additionally, we provide a demonstration of the tool's use and utility by conducting data simulations and a case study testing real hypotheses.

Chapter 2

Effect size

> "*[...] the emphasis given to formal tests of significance [...] has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, [...] and too little to the estimates of the magnitude of the effects they are investigating.*"
>
> — Yates, 1951, p. 32

In this chapter, we focus on the concept of effect size. To begin, we consider it essential to clarify the approach we have chosen for discussing hypothesis testing, as well as the distinction between statistical and practical significance. In line with our objective of concentrating on the latter, we present the definition of effect size and outline various methods for its calculation. Given that Cohen's d is one of the most widely used effect size indices, particularly in psychological research, we center our attention on its interpretation, limitations and potential improvements.

2.1 Overview

Although this is not the most common occurrence in psychological research, much applied research begins with a research hypothesis that states that there is a relationship

between two variables or a difference between two parameters, such as means (in subsequent chapters, we explore research that involves more than two variables). One typical form of the research hypothesis concerns a possible relationship between the two variables in the population. Often one variable is a categorical independent variable involving group membership (a grouping variable), such as male versus female or Treatment a versus Treatment b, and the other variable is a continuous dependent variable, such as blood pressure or score on an attitude scale or on a test of mental health or achievement (Rosenthal et al., 2000).

2.1.1 Hypothesis testing in social sciences

The two dominant inferential approaches in the social sciences are the frequentist approach and the Bayesian approach. For a more in-depth explanation of the latter, refer to the article by Van de Schoot et al. (2014). Both schools of thought are valid; however, as they propose different ways of approaching research, in this work, we will follow the one that best aligns with our objectives, namely the frequentist approach.

A frequentist can be defined as a researcher who approaches issues of probability in terms of the frequency (number of occurrences) for a particular parameter over a period of time (VandenBos, 2007). Within a frequentist framework, a researcher can adopt one of many approaches to test hypotheses. For example, Fisher's null hypothesis testing can be applied by establishing a statistical null hypothesis, where the null hypothesis does not necessarily need be a nil hypothesis (i.e., zero difference), and then reporting the exact level of significance (e.g., $p = 0.051$ or $p = 0.049$) (Fisher, 1955). According to Fisher, null hypothesis testing represents the most rudimentary form of statistical analysis and should be employed only in contexts where we possess little or no prior knowledge (Gigerenzer et al., 1990).

Neyman and Pearson criticized Fisher's null hypothesis testing for several reasons, including that no alternative hypothesis is specified. In its simplest version, Neyman–

Pearson theory has two hypotheses and a binary decision criterion (Neyman, 1957). They indeed suggested to establish two statistical hypotheses, H_0 and H_1, and determine the values of α (the probability of committing a Type I error or incorrectly rejecting the null hypothesis when it is true) and β (the probability of committing a Type II error or failing to reject the null hypothesis when it is false) as well as the sample size, prior to conducting the experiment. These decisions should be guided by subjective cost-benefit analyses. According to their approach, if the data fall within the rejection region of H_0, H_1 should be accepted; otherwise accept H_1. The applicability of this procedure is primarily limited to scenarios where there is a clear disjunction of hypotheses and where it is possible to make meaningful cost-benefit trade-offs when selecting α and β. In the Neyman-Pearson approach, the power of a statistical test is defined as the probability that the test has to reject the null hypothesis (H_0) when the alternative hypothesis (H_1) is true (Altoè et al., 2020).

The misguided guidelines presented in textbooks and manuals that have shaped generations of psychologists and researchers gave rise to a third approach, known as *null hypothesis significance testing* (NHST), an inconsistent hybrid of the two previously mentioned theories: Fisher's null hypothesis testing and Neyman and Pearson's decision theory (Gigerenzer, 2004). This approach to hypothesis testing is neither practiced in the natural sciences nor in the field of statistics proper; however, it has become institutionalized and widely adopted in psychology and many other scientific disciplines. According to Gigerenzer (2004), who refers to it as the 'null ritual,' NHST has become a source of confusion within these fields. The null ritual comprises three steps: establish a statistical null hypothesis without specifying either your own hypothesis or an alternative hypothesis, apply a 5% significance level to reject the null hypothesis and accept your own hypothesis and, finally, always adhere to this procedure (Gigerenzer, 2004). The primary weaknesses of NHST include: first, the absence of a clearly defined research hypothesis; second, the lack of power calculations and, consequently, a lack

of pre-planned sample size; and third, the suboptimal default use of an alpha level of 0.05, without justifying a more appropriate alpha level (Maier & Lakens, 2022), which leads to the mistaken belief that rejecting H_0 automatically supports the researcher's hypothesis, often not formally defined.

To ensure clarity and avoid the confusion surrounding testing approaches highlighted by Gigerenzer (2004), we begin our discussion by explicitly stating the chosen approach, thereby engaging in rigorous statistical thinking. The Neyman–Pearson testing framework will serve as the foundation for the concepts discussed in the following sections, as well as for the explanation and application of the proposed framework.

2.1.2 Statistical and practical significance

As previously mentioned, for more than a century, researchers have stressed the importance of not just interpreting the statistical significance of effects, but also their practical significance (Yates, 1951). A statistically significant result is signifying that the result is sufficient, by the researcher's adopted standard of required evidence against the hypothesis H_1 (say, adopted significance level $\alpha < 0.05$), to justify rejecting H_0 (Grissom & Kim, 2012). For instance, the difference between the experimental group and the control group as statistically significant represents relevant information. However, this observation does not provide any insight into the magnitude of the difference. A difference that is statistically significantly different from zero may have little to no practical or clinical relevance when considering the estimated effect size between the two groups. A psychotherapeutic example of a statistically significant result that would not be practically significant in the sense of *clinical significance* would be a statistically significant lowering of scores on a test of depression that is insufficient to be reflected in the clients' behaviors or self-reports of well-being. Effect size indices are utilized to assess the magnitude of the effect of interest, irrespective of the measurement units of the variables under investigation. In conjunction with the inferential results

of an analysis, the effect size should be carefully evaluated by the researcher to ensure a proper interpretation of the phenomenon under study. Although effect size is not synonymous with practical significance, knowledge of a result's effect size can inform a subjective judgment about practical significance, which is enhanced by expertise in the area of research (Grissom & Kim, 2012). It is therefore evident how crucial statistical inference is within the research overview, but both in general and, in particular, in social sciences practical significance of effects must be taken into consideration as an important component of a study, which is instead turns out to a process often neglected, as reported above (Dienes, 2008).

2.1.3 Definition of effect size

To provide a precise definition of the concept of effect size, we consider the case of a typical hypothesis that implies there is no effect or no relationship between variables. Whereas a test of statistical significance is traditionally used to provide evidence (attained p level) that the hypothesis is wrong, an effect size measures the degree to which such hypothesis is wrong (if it is wrong) (Grissom & Kim, 2012). Researchers are often reminded to report effect sizes because of what has been said so far about the matter of practical significance, but also for two other reasons: firstly, effect sizes on one hand allow researchers to draw conclusions by comparing standardized effect sizes across studies, secondly effect sizes from previous studies can be used when planning a new study in an a-priori power analysis (Lakens, 2022). To underscore the methodological significance of effect sizes, the following is a direct statement from the American Psychological Association (APA) on this matter: "[Significance testing hypothesis] is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results." (American Psychological Association, 2010, p. 33).

2.2 Effect sizes based on means

When studies report means and standard deviations, two commonly preferred effect sizes are the raw mean difference and the standardized mean difference (Borenstein et al., 2021). In this section, these two effect size indices will be discussed separately.

2.2.1 Raw mean difference

For raw (unstandardized) mean difference, the effect size is expressed on the scale on which the measure was collected. The primary advantage of the raw effect size is that it is intuitively meaningful, either inherently (for example, blood pressure, which is measured on a known scale) or because of widespread use (for example, a national achievement test for students, where all relevant parties are familiar with the scale) (Borenstein et al., 2021). In a study that reports the means for two groups (treated and control, denoted as μ_1 and μ_2), where a comparison of these group means is required, the population mean difference is defined as

$$\Delta = \mu_1 - \mu_2.$$

It is possible to perform a meta-analysis or a power analysis based on unstandardized effect sizes and their standard deviation, but it is easier to work with standardized effect sizes, especially when there is variation in the measures researchers use. To facilitate a comparison of effect sizes across situations where different measurement scales are used, researchers can report standardized effect sizes (Lakens, 2022). This is way more useful in psychology where the studied phenomena are assessed with different scales or so they need to be standardized in order to be compared and allow researchers to draw useful conclusions.

2.2.2 Standardized mean difference

The raw difference between the means of two groups depends on the unit of measurement, however, in fields such as psychology, it is more common to employ standardized measures of effect size. This preference arises because the variables in this field often lack specific units of measurement (e.g., anxiety, empathy, memory, or learning), or because different instruments are used to measure the same phenomenon (e.g., various psychological tests) (Borenstein et al., 2021). Furthermore, unstandardized effect sizes do not account for the variability of the quantitative variable within the two groups (Baguley, 2009). To address these concerns, we can refer to one of the most widely recognized standardized indices for measuring effect size, namely Cohen's d.

Cohen's d is a standardized measure of effect size that expresses differences in terms of the variability of the phenomenon under investigation, independent of the original unit of measurement. It provides a valuable solution when researchers employ raw units that are either arbitrary or lack meaningful interpretation beyond the specific context of their study (Cohen, 1988). Cohen's d (δ) is defined as the raw difference between two population means ($\mu_a - \mu_b$) divided by the common standard deviation (σ):

$$\delta = \frac{\mu_a - \mu_b}{\sigma}$$

For instance, A Cohen's d of 0.1 means that the difference between the two population means is one-tenth of the common standard deviation. Notably, this index is independent of the unit of measurement of the phenomenon under investigation (Altoè et al., 2020). Borenstein et al. (2021) underline the importance of distinguishing between δ, the population Cohen's d value, and d, the estimated Cohen's d value from the sampled groups given by:

$$d = \frac{\bar{X}_a - \bar{X}_b}{S_{\text{pooled}}}$$

where

- S_{pooled} represents the pooled standard deviation of the two samples

$$S_{\text{pooled}} = \sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}};$$

- $\bar{X}_a - \bar{X}_b$ represents the difference between the means of groups a and b;
- s_a^2 and s_b^2 represent the the variances of groups;
- n_a and n_b indicate the sample sizes of the groups.

We present a graphical representation of the concepts discussed thus far. An example of a meta-analytic effect size is the finding that individuals working in a group exert less effort to achieve a goal compared to those working individually, a phenomenon known as *social loafing*, which is quantified with an effect size of $d = .43$, an effect is substantial enough to be noticeable in everyday life (Lakens, 2022). However, when examining the overlap between the two distributions (Cohen's U_3), it becomes evident that there is considerable overlap in the effort exerted by individuals across the two conditions (individual work versus group work). As illustrated in figure 2.1, the probability of superiority - or the likelihood that a randomly selected individual from the group condition exerts less effort than one from the individual condition - is only 61.9%.

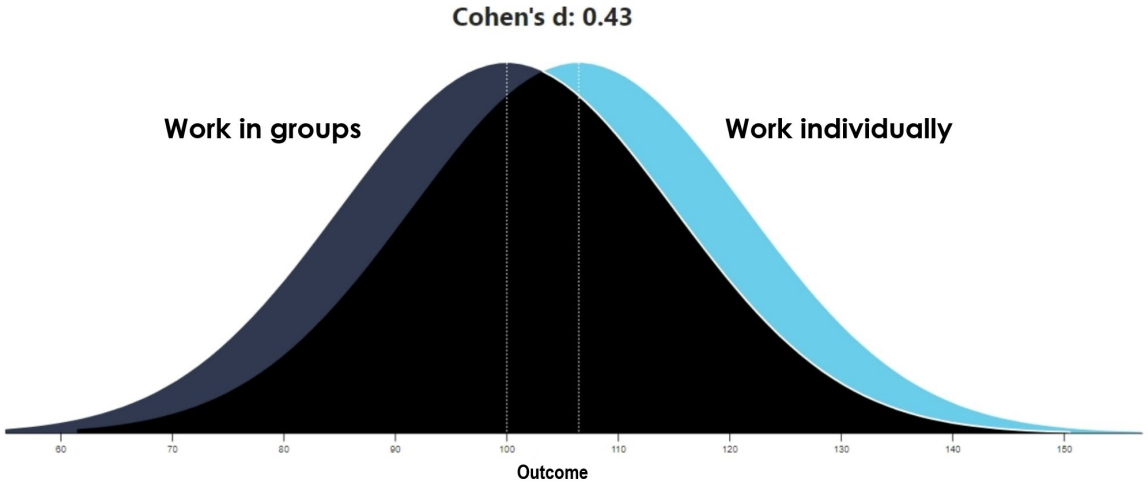

Figure 2.1: A visualization of 2 groups representing $d = 0.43$ from Magnusson, K. (2023), A causal inference perspective on therapist effects

One contributing factor to the widespread use of Cohen's d is the general acceptance of established benchmarks for interpreting this effect size (McGrath & Meyer, 2006). These benchmarks, along with related considerations, will be further examined in the following section.

2.2.2.1 Unbiased effect size estimate

Hedges (1981) demonstrated that d exhibits a slight bias, tending to overestimate the absolute value of δ in small samples. This bias can be corrected through a simple adjustment, resulting in an unbiased estimate of δ, referred to as Hedges' g. To convert d to Hedges' g, a correction factor, known as J, is applied (Borenstein et al., 2021). While Hedges (1981) provides the exact formula for J, in practice, researchers commonly use an approximation

$$J = 1 - \frac{3}{4df - 1}$$

In this expression, df is the degrees of freedom used to estimate S_{pooled}, which for two independent groups is $n_a + n_b - 2$. Then,

$$g = J \times d$$

The correction factor (J) is always less than 1.0, resulting in g being smaller than d in absolute terms. Consequently, when the sample size is large and the degrees of freedom (df) are high, J approaches 1.0 and the difference between d and g becomes negligible. Conversely, when the degrees of freedom are very small (e.g., fewer than 10), the difference between these two estimates of δ - the population effect size - becomes more pronounced.

2.3 Interpreting effect sizes

Merely reporting an effect size without properly interpreting it adds little to a report of research. The American Educational Research Association recommends including an estimate and interpretation of effect size for each important inferential statistic that is reported (Grissom & Kim, 2012). Unfortunately, a considerable amount of work remains to be done in this area. Many researchers continue to adhere to outdated guidelines, often due to a misinterpretation of the original recommendations provided in the past.

2.3.1 Cohen's benchmarks

Typically, effect sizes are stated, but a repeatedly noted problem is that effect sizes are either not interpreted at all or are at best 'labeled' according to J. Cohen's benchmark values. Cohen (1988) proposed $d = 0.2$, $d = 0.5$, and $d = 0.8$ as small, medium, and large effect sizes respectively in the field of social sciences. Based on his proposal, an effect size of $d = 0.3$ indicates a difference in means that is difficult to detect (e.g., the average height difference between 15- and 16-year-old girls). Moreover, an effect size of $d = 0.8$ refers to a completely obvious difference in means, such as the average height difference between 13- and 18-year-old girls. The medium effect size of $d = 0.5$ was defined as "large enough to be visible to the naked eye" (p. 26). Even though Cohen 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., 2023).

Cohen (1988) reluctantly used these conventions in the context of power analysis

"only when no better basis ... [was] available" (p. 25) and later told friends that he actually regretted having suggested them at all. His regret was indeed well-founded, since these standards are meaningless in the absence of a frame of reference (Funder & Ozer, 2019).

To not just report but interpret an effect size, nothing is gained by the common practice of finding the corresponding verbal label of 'small', 'medium' or 'large' (Lakens, 2022). In the next section will be presented some recent approaches suggested recently to address the matter of meaningful interpreting effect sizes.

2.3.2 Recent approaches

For decades, researchers have cautioned against the use of Cohen's benchmarks (Correll et al., 2020) and a critical advancement has been the recognition of the issues associated with these benchmarks. For instance, Funder & Ozer (2019) highlighted the widespread disregard for effect size within the academic careers of many psychologists, demonstrating that reliance on Cohen's standards often leads to uninformative and misleading interpretations. Although their intentions are aimed at developing meaningful interpretations, Funder & Ozer (2019) ultimately recommended comparing the magnitude of a finding with another well-understood finding. However, this approach does not fully move away from the use of benchmarks.

Claims based on data become meaningful when observed effects are related to their theoretical or practical consequences, but the dominant approach to statistical inferences in psychology, null hypothesis significance testing (NHST), represents a limit to it. One of the most widely suggested improvements in the use of p-values is to replace null hypothesis tests, in which the goal is to reject an effect of exactly zero, for tests of range predictions, in which the goal is to reject effects that fall outside the range of effects that is predicted (Lakens, 2021). One approach to achieve this is to use the two one-sided tests (TOST) procedure, which tests for equivalence rather than against zero,

allowing researchers to reject the presence of a smallest effect size of interest (SESOI) (Lakens et al., 2018). Another example is provided by Hilgard (2021), who suggests that a prudent practice is to estimate the largest plausible effect size - defined as the upper limit of plausible effect size for a given measure - and subsequently test this estimate.

By adhering to these procedures, psychologists would be able to test not only for the presence of an effect but also for the absence of an effect that is either too large or too small to be real or practically useful (Lakens, 2021). However, further progress is needed in understanding how an effect translates to the dependent variable: current methodologies, including newer approaches, still often treat effect size merely as a numerical value. Researchers must explicitly identify the relevant mechanisms that represent the effect in the context to which they are generalizing (Anvari et al., 2023), given the urgent need to advance beyond merely falsifying a numerical value, as relevant as that improvement may be, and to focus on falsifying the underlying model of the effect.

2.4 Limitations of Cohen's *d*

One of the primary objectives of statistical analysis is to examine the relationship between variables. An association exists between two variables when the values of one variable tend to vary systematically with changes in the other (Bertani et al., 2018). For instance, in psychological research, a common example of an association might involve a variable called "Response to treatment" and another variable labeled "Treatment", which could take on values such as "Treatment *a*" and "Treatment *b*." If "Treatment *b*" represents a placebo, it is likely that individuals receiving "Treatment *a*" will show greater improvement compared to those receiving "Treatment *b*." In this scenario, an association is observed between the variables "Response to treatment" and "Treatment,"

as the proportion of individuals responding to treatment varies according to the type of treatment administered. Typically, when analyzing the association between two variables, a process known as "bivariate analysis", one variable is designated as the "outcome variable" and its different values are compared based on the values of the other variable, referred to as the "explanatory variable". In other words, bivariate analysis facilitates the evaluation of how the value of the outcome variable is influenced by, or can be explained by, the values of the explanatory variable (Agresti, 2012).

As shown in the second section of this chapter, the magnitude of the difference between the mean outcome values of two groups defined by the explanatory variable is referred to as the effect size (Coe, 2002). Cohen's d is well-suited for application in such contexts: these contexts are bivariate, since they concern only the explanatory and the explanatory variable, and present the particular feature of the explanatory variable as categorical with two levels and the outcome variable as continuous. For instance, when comparing gender and income, income serves as the outcome variable, while gender (male versus female) is the explanatory variable, and the effect size represents how much being male or female influences the mean income of these two groups. In contrast to analysis of data from two variables (bivariate), multivariate analysis is concerned with a group (or several groups) of individuals, each of whom possess values or scores on two or more variables, such as tests or other measures (Tatsuoka & Lohnes, 1988). Multivariate data can provide a more comprehensive understanding of a sample population than bivariate data. In fact, many of the most compelling research questions in the social sciences are inherently multivariate in nature (Hammond, 2000). Furthermore, as is frequently the case in both the biomedical and social sciences, many quantitative studies incorporate qualitative variables. These variables, also referred to as categorical variables, often consist of more than two categories. They are known as multinomial, representing the various forms in which a variable may differ across contexts (e.g., sexual orientation, ABO blood type, marital statusA or religion). This highlights the inherent

complexity of the phenomena examined by these disciplines (Miola & Miot, 2022).

This brings us to a notable limitation of Cohen's d: because it quantifies the difference between the means of two groups, it proves challenging to apply in contexts both multivariate and concerning multinomial independent variables. Given that such contexts are common in fields that study complex phenomena and processes, such as psychology, addressing this limitation is of considerable importance. Earlier in this chapter, we discussed the importance of designing studies aimed at falsifying models that represent effect sizes. This approach would allow us to address two key objectives relative to hypothesis testing: aligning theoretical models with appropriate statistical methods and focusing on models of effect sizes rather than on isolated values. Such an approach would contribute to more meaningful research in the social sciences, which have been plagued by both theoretical and replicability crises.

Furthermore, it can also be stated there is a pressing need to develop methodologies that enable the use of Cohen's d, one of the most widely utilized effect size indices, within the context of multivariate analyses, given their prevalence in psychology. In conclusion, advancements are necessary in the areas of falsifying both effect size models and the meaningful application and interpretation of Cohen's d. Such improvements would greatly benefit the field of psychology, where Cohen's d is among the most widely used effect size indices, and where a deep understanding of how effects, such as treatments or drugs, operate is of fundamental importance.

Chapter 3

Proposed framework

> *"Thinking hard about effect sizes is important for any school of statistical inference [i.e., Frequentist or Bayesian], but sadly a process often neglected."*
>
> — Dienes, 2008, p. 92

In this chapter, we outline the key characteristics and features that distinguish widespread models in the representation of psychological phenomena, specifically those involving moderator and mediator variables. After clarifying the distinctions between full and partial mediation, we present a practical example of a full mediation model drawn from the literature. This type of mediation will serve as the basis for the concrete application of the proposed framework in subsequent sections.

3.1 Moderator and mediator variables in psychological research

Given the multifaceted and complex nature of phenomena and processes studied in psychology, multiple variables often contribute to outcomes of interest. Consequently, researchers have access to statistical methods specifically designed to address the relationships among multiple variables and outcomes. This distinction is particularly

relevant in psychology (Ballen & Salehi, 2021). Such models are invaluable from the early stages of research, particularly when designing a study. After selecting the variables to manipulate and those to measure within a specific subject, researchers must carefully consider the relationships among these variables, as well as the potential influence of third variables that may affect the relationship under investigation, acting as moderator or mediator variables. Although the concepts of moderator and mediator variables have a relatively long-standing tradition in the social sciences, it is not uncommon for social psychological researchers to use these terms interchangeably (Baron & Kenny, 1986).

A possible function that third variables can cover is acting as **moderator variables**, affecting the size or nature of the relationship between an independent and dependent variable. (Koeske, 1992). Following the approach of Baron & Kenny (1986), the essential properties of a moderator variable can be illustrated through the findings of Glass & Singer (1972). Glass and Singer hypothesized that an interaction of the factors *stressor intensity* (noise level) and *controllability* (periodic-aperiodic noise) of the form that an adverse impact on task performance occurred only when the onset of the noise was aperiodic or unsignaled. The essential properties of such a moderator variable model are summarized in figure 3.1. The model diagrammed in figure 3.1 presents three causal paths that feed into the outcome variable of task performance: the impact of the noise intensity as a predictor (path a), the impact of *controllability* as a moderator (path b) and the interaction or product of these two (path c). The moderator hypothesis is supported if the interaction (path c) is significant.

On the other hand, the **mediator function** of a third variable represents the generative mechanism through which the independent variable is able to influence the dependent variable of interest (Baron & Kenny, 1986). The outcome can be affected by an independent variable both indirectly through a mediating variable (path encompassing arrows a and c) and directly (path of arrow c), as shown in figure 3.2.

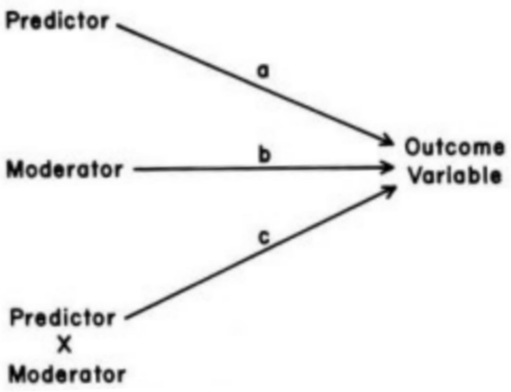

Figure 3.1: Moderator model by Baron, R. M. and Kenny, D. A. (1986), The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, Journal of personality and social psychology, p. 1174

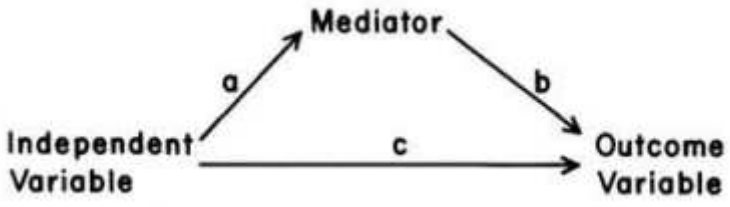

Figure 3.2: Mediation model by Baron, R. M. and Kenny, D. A. (1986), The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, Journal of personality and social psychology, 56(6), p. 1176

Two forms of mediation models can be distinguished based on the different ways in which phenomena operate and occur. In addition to identifying mediators and mediation pathways, finalizing the mediation structure requires determining whether a partial or full mediation model is more appropriate for the analysis, as illustrated in figure 3.3 (Ballen & Salehi, 2021). These two types of mediation models, which reflect the different pathways connecting the independent variable to the dependent variable, are referred to as partial or full mediation models. Ballen & Salehi (2021) describe the contrast between partial and full mediation models to test the mediating effect of variable B in the relationship between variable A and variable C. The **partial mediation model** (top) tests the partial mediation effect of A on C. In this model, A directly affects C (i.e., A → C) and indirectly via B (i.e., A → B → C). The **full mediation model** (bottom) tests how B fully mediates the relationship between A and C (i.e., only A → B → C). The encircled "e" acknowledges error inherent to model estimates. Variables can be either observed (rectangles) or latent (ellipses).

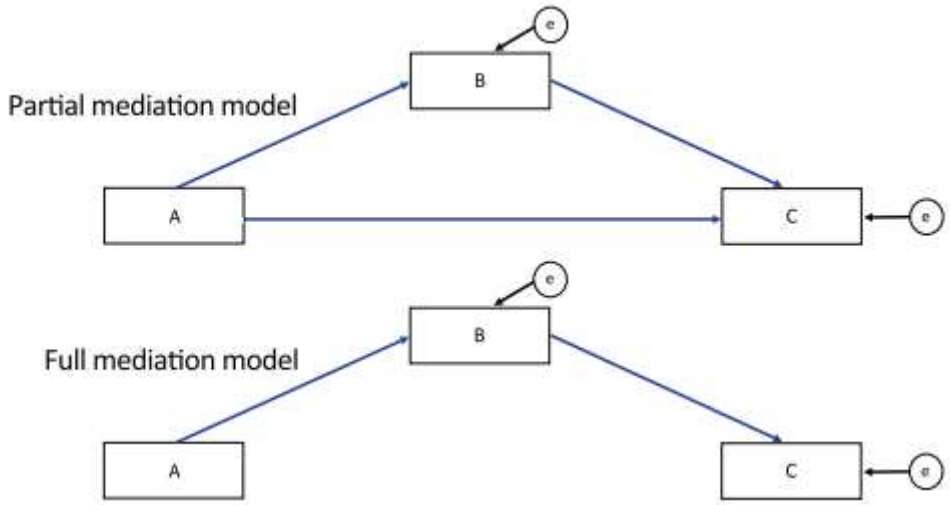

Figure 3.3: Contrasting a partial and full mediation model by Ballen, C. J. and Salehi, S. (2021), Mediation analysis in discipline-based education research using structural equation modeling: Beyond "what works" to understand how it works, and for whom, Journal of microbiology and biology education, (22,2) p. 5

For instance, Ballen & Salehi (2021) proposed that mediation analysis can be employed

to examine the mediating impact of a student's prosocial utility value beliefs in biology (e.g., using biology to achieve goals related to helping others) as a mediator for the effect of using a textbook with prosocial examples (or a neutral control, as the independent variable) on the student's interest in the subject of biology (the dependent variable) (Zambrano et al., 2020). They hypothesized that the textbook condition (prosocial or control) influences interest through prosocial utility value beliefs. Specifically, using a textbook with prosocial examples is expected to enhance students' beliefs in the social utility value of biology, which, in turn, increases their interest in the subject. The extent to which students' prosocial utility value beliefs in biology mediate the effect of textbook condition on their interest in the subject can be quantitatively assessed. If the mediation effect has a non-significant p-value, or a significant p-value with a small estimated effect, one may conclude that the effect of textbook condition on student interest in biology is primarily mediated through mechanisms other than the student's prosocial utility value beliefs. Conversely, if the size of mediation is large, it suggests that a student's prosocial utility value beliefs are a significant mediator in the relationship between textbook condition and student interest. In such cases, reading a textbook with prosocial examples may lead to stronger prosocial utility value beliefs, which, in turn, enhances interest in biology. If, after accounting for the mediating effect of prosocial utility value beliefs in biology, the textbook condition still exerts a significant direct effect on student interest, this would indicate an example of a partial mediation model. Otherwise, it would suggest a full mediation model.

Psychologists have long recognized the significance of mediating variables (Baron & Kenny, 1986). Consequently, our focus will shift to mediation models due to their critical importance and their involvement in numerous psychological theories, both historical and contemporary. Additionally, mediation models are relatively straightforward to understand. Given that our aim is to introduce a framework incorporating novel elements, we believe the most effective approach is to demonstrate this framework through

examples applied in the simplest scenarios.

3.2 A full-mediator model within the proposed framework

To resume the discussion from the conclusion of the second chapter, we addressed the recommendation of designing studies based on hypotheses about models of effect size that can be falsified by data as a potential solution to the replication crisis in psychology. In line with this, we began this chapter by discussing models, as the primary aim of this study is to present a framework of falsifiable effect size models and to provide a concrete application, demonstrated through data simulation and the analysis of a real dataset.

When selecting the scenario in which to implement this framework, we determined that it should involve a model incorporating at least one third variable. This decision reflects the fact that psychological processes rarely involve a simple relationship where an independent variable directly influences a dependent variable, since outcomes are more often influenced by additional variables that play a role in the mechanism under investigation. Furthermore, at the end of the second chapter, we highlighted the importance of addressing the use of Cohen's d as a measure of effect size within the context of multivariate analysis. In this regard, we opted for a full mediation model, as it represents a simpler and more straightforward scenario compared to partial mediation. Given that our goal is to introduce a novel framework, we believe that simplicity is the most effective approach for explanation. For this reason, we chose to apply the proposed framework within a multivariate analysis involving only three variables.

This model possesses two key features that are crucial to consider, as they define the specific scenario we aim to explore through simulation and data analysis. By focusing on the full mediation model, we restrict our example to contexts in which the entire effect of the manipulation is transferred through the mediator, thereby assuming no

direct impact on the dependent variable. While this represents a particular case, there are instances in which researchers specifically hypothesize such relationships between variables, as will be discussed later. In this model, the manipulation of the independent variable directly alters the mediator, which in turn influences the dependent variable. The second important feature is the causal relationship between the variables: the manipulation must initiate a chain of events and the mediator must be a variable capable of change. For example, variables such as anxiety levels or reaction times are suitable mediators, whereas static variables like sex or age are not.

The simplicity of the scenario chosen for the application of this framework serves as a foundation for future developments. Once researchers have grasped its functionality in this basic context, they can extend its application to more complex situations, involving more than three variables that may interact in different ways, such as through partial mediation or moderation. Additionally, while this scenario is relatively simple, it reflects a structure that may be genuinely hypothesized by researchers. In the following section, we will illustrate the full mediation model through a concrete hypothesis, as implemented in the study by Reinhold et al. (2018). Before presenting the example, it is important to emphasize that the research discussed subsequently involves a more complex case than the scenario for which we intend to apply our framework. Specifically, Reinhold et al. (2018) addresses a predictor variable defined by four dimensions (rather than a simple, single-dimensional variable), as will be clarified in greater detail later.

3.2.1 Example of a full mediation model from literature

In their study, Reinhold et al. (2018) examine the concept of transfer of training, which they define as the application of acquired knowledge and skills in the workplace. Previous research has identified social support and motivation to transfer as two key factors in explaining this process. Furthermore, as the authors note, certain models

delineate more specific sub-dimensions of social support, including supervisor support, peer support, supervisor sanctions and feedback/coaching. However, it remains unclear whether social support directly influences transfer of training or if this effect is mediated by motivation to transfer. The researchers aim to understand the relative impact of these sub-dimensions of social support on both motivation to transfer and the transfer of training itself. By testing both partial and full mediation models, their findings suggest that the sub-dimensions of social support exert an indirect influence on transfer of training through motivation to transfer, with the full mediation model providing a better fit (Reinhold et al., 2018). In this context, we focus on the full mediation model tested in their study to gain a clearer understanding of the role of mediating variables in psychological research. As one of the initial hypotheses proposed, and later confirmed as the better explanation, motivation to transfer is a crucial element in understanding the relationship between learning and behavior change. As illustrated in figure 3.4, the effect of all predictor variables is fully mediated by the mediator, indicating that social support does not have a direct influence on transfer. Instead, the influence of all predictor variables is mediated by transfer motivation.

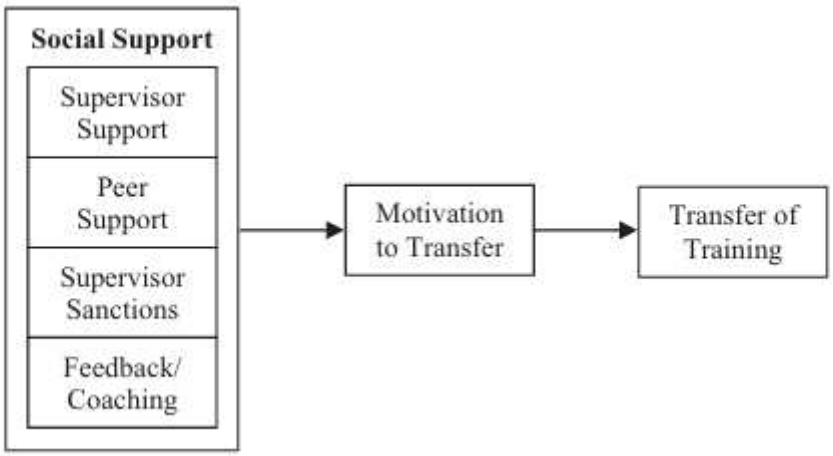

Figure 3.4: A fully mediated model of the hypothesized relationships by Reinhold, S., Gegenfurtner, A. and Lewalter, D. (2018). Social support and motivation to transfer as predictors of training transfer: Testing full and partial mediation using meta-analytic structural equation modelling. International Journal of Training and Development, 22(1), p. 3

In conclusion, the next chapter will apply the proposed framework within the context of a full mediation model involving three variables, similar to the example just discussed. Although this setting may appear simple and well-defined, it nonetheless represents a viable research question, and in certain cases - such as the one just considered - proves to be the best fit for explaining psychological mechanisms. Consequently, the demonstration of how to implement the proposed framework in this context is not only valuable from a didactic and theoretical perspective but can also be immediately applicable in psychological research practice.

3.3 Implementing the framework in research

The objective of this dissertation is to address several flawed practices that ultimately undermine the reliability of scientific research by proposing improvements that can be implemented in the conduct of confirmatory research. To achieve this, we propose a framework that emphasizes the meaningful interpretation of effect sizes, particularly Cohen's d, within the context of multivariate analysis, supported by data simulations and graphical representations of predicted models.

In particular, we suggest that researchers can initiate this improvement process by formulating hypotheses related to the effect size of interest and carefully reasoning about how the effect of the manipulated variable translates to the outcome. This approach enables the development and testing of statistical models that align with theoretical assumptions. At this stage, it is crucial to simulate the expected results and, within these simulations, calculate the study's statistical power a priori. Moreover, we propose that graphical representations in this context can provide researchers with a clearer and more intuitive understanding of the hypotheses being tested by visualizing simulated outcomes through plots. Once the registered report has undergone peer review, researchers can proceed with data collection and analysis, ultimately plotting

and visualizing the final results. At this stage, they can compare both the numerical and graphical outputs from the simulations and the real data analysis, leading to more meaningful conclusions as part of the falsification process.

By incorporating both simulations and graphical representations in the study design, as well as comparing simulated and real data results, we aim to achieve the following objectives: first, encouraging researchers to consider the magnitude of the effect size from the outset of the study, fostering a more thoughtful approach to hypothesis formulation; second, promoting a more meaningful interpretation of effect size, moving beyond its treatment as a mere numerical value calculated at the conclusion of the analysis.

Chapter 4

Data simulation

> *"We can go further by considering what comes before data analysis: design of experiments and data collection. [...] We can often do better by preceding any data collection with a computer simulation."*
>
> — Gelman, 2023, p. 1-2

This chapter will be dedicated to data simulation. We will begin by explaining how simulation can contribute to enhancing the rigor and effectiveness of confirmatory research, which is a key focus of this dissertation. Following this, we will outline the full mediation model that forms the basis of our simulation, using a concrete psychological example to illustrate its application. Subsequently, we will introduce the multivariate simulation, detailing its characteristics and the procedural steps implemented within the R script. Lastly, we will discuss how the proposed simulation and visualization framework can be effectively applied by researchers in practice.

4.1 Illustration of the employed full-mediation model

4.1.1 Importance of simulating

In previous chapters, we have discussed the significance of each step involved in experimental design, particularly those preceding data analysis. One crucial step that warrants further emphasis is the use of simulations and preliminary analyses before data collection. Data simulation is a powerful tool for studying complex systems and predicting their behavior, as it allows for the controlled reproduction of a wide range of scenarios, enabling the prediction of outcomes and testing of various models and hypotheses. As noted by Gelman (2024), beyond the evident benefit of identifying designs that may be too noisy to detect main effects or interactions, constructing simulations sharpens researchers' focus by necessitating deliberate decisions about the structure and magnitude of effects. Through simulations, researchers can make informed assumptions about measurement variations and treatment effects, thereby facilitating the use of metrics that accurately reflect the underlying constructs of interest. This, in turn, allows for more realistic effect size estimations, which are crucial for enhancing the validity of confirmatory research.

To encourage the adoption of this approach, we aim to present simulations of the aforementioned full-mediation model for educational purposes. The primary objective of this chapter is to emphasize the importance of thoughtful planning before empirical testing, with the goal of aligning theoretical models with appropriate statistical methods, ultimately leading to more accurate interpretations of effect sizes. To achieve this, we propose a method for visualizing the magnitude of outcome effects through graphical representations, accompanied by the necessary R code. This approach allows for the direct simulation of various scenarios, which are commonly encountered during the initial phases of study design. Additionally, it underscores the critical role of simulations in refining experimental frameworks.

4.1.2 Employed full-mediation model

In this chapter, we aim to conduct a multivariate simulation, as we are simulating more than one variable — specifically two variables — simultaneously. It is important to note that we are considering a trivariate function, since our model involves an independent variable, a dependent (outcome) variable and a mediator. To clarify the context in which our framework will be applied through data simulation, we will reference the study conducted by Zambrano et al. (2020), as it provides a suitable model for the setting we have chosen. Their investigation serves as an appropriate example for illustrating the interaction between the variables of interest within a full-mediation model, aligning with the approach we have adopted. Additionally, Ballen & Salehi (2021) utilized the same study to demonstrate the distinction between full and partial mediation models, as discussed in Chapter 3. In light of this, we now present the example, applying the full-mediation model described earlier to facilitate our discussion.

As illustrated in figure 4.1, the study examines whether incorporating prosocial utility value into a science textbook chapter enhances students' perceptions that the chapter's topic provides opportunities to engage in prosocial activities and increases their interest in the subject. Specifically, as suggested by one of the hypotheses from Zambrano et al. (2020), it is anticipated that students who read a textbook containing prosocial information (compared to those reading control textbooks) will be more likely to report that the science topic offers opportunities to achieve prosocial goals. This contrasts with students who read a chapter that does not emphasize utility value. The total sample of the study consists of all participating students, who are randomly assigned to one of two groups: the experimental group (prosocial utility value condition) and the control group. In the experimental group, students read a textbook chapter enriched with examples or connections illustrating how the topic or concept can be applied to benefit humans. In contrast, the control group reads a textbook chapter

without any additional utility value content.

As discussed in previous chapters, this dissertation aims to facilitate more meaningful interpretations of effect sizes, particularly those measured by Cohen's d, through the use of simulations and visual tools, especially in the context of models involving multiple variables, as is common in psychological research. In applying this objective to the present multivariate example, we focus on modeling and visualizing how the effect of prosocial textbooks on prosocial affordance beliefs subsequently influences the interest towards scientific topics. We will pursue this goal by first formalizing the statistical model representing the full-mediation relationship, chosen as the simplest framework for implementing our approach. This will be followed by the simulation of various scenarios to illustrate its application.

Figure 4.1: Full mediation model of textbook condition predicting interest through prosocial affordance beliefs adapted from Zambrano, J., Lee, G. A., Leal, C. C. and Thoman, D. B. (2020). Highlighting prosocial affordances of science in textbooks to promote science interest. CBE - Life Sciences Education, 19(3), ar24, p. 7

4.2 Multivariate simulation

4.2.1 Formalization of the statistical model

The multivariate framework in which we intend to implement our approach, as articulated in the previously described concrete example, is delineated as follows: the

predictor (prosocial textbook, denoted as X) is a dichotomous categorical variable that exerts an influence on the outcome (science topic interest, denoted as Y), which is characterized as a continuous variable. The effect of the predictor is fully mediated by prosocial affordance beliefs, another continuous variable that functions as the mediator, hereinafter referred to as M. One foundational assumption underpinning the hypothesis of this proposed study is the absence of a direct effect of textbook condition on interest. Additionally, it is crucial to assume that the two samples represented by the dichotomous values of X - the experimental group (where the students read a textbook chapter enriched with prosocial utility value) and the control group (where the students read a textbook chapter without any additional utility value content) - are independent.

In the context of this trivariate function, the two samples are assumed to follow multivariate normal distributions, centered around the mean values observed for the two variables: "prosocial affordance beliefs" and "science topic interest". For the experimental group, d_1 denotes the difference in "prosocial affordance beliefs" scores measured before and after the exposure to the prosocial utility value condition. Conversely, d_2 reflects the corresponding difference for the "science topic interest" variable. Since the control group is not exposed to the experimental condition, we assume that both d_1 and d_2 are equal to 0 for subjects within this group. These variables are assumed to be standardized within each group, with a variance of $\sigma^2 = 1$, and exhibit as correlation coefficient r. Since the variables are standardized, the covariance values are equivalent to the correlation values.

Consequently, Y is characterized as a multivariate variable that follows a normal distribution centered on d, which is a vector representing the mean values of the pre- and post-exposure differences. This variable is associated with a covariance of r and a variance of 1. Within each group, this relationship can be mathematically articulated through the following equation:

$$\mathbf{Y} \sim \mathcal{N}(d, \Sigma)$$

where Σ represents the variance-covariance matrix.

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

4.2.2 Features of the simulation

One notable simulation technique is the Monte Carlo method, which is extensively utilized across various scientific disciplines, particularly in the field of statistics. The fundamental principle underlying Monte Carlo simulation involves generating a series of simulated scenarios, each characterized by a unique set of parameters. For each scenario simulated, data are generated randomly in accordance with the parameters of interest to observe their behavior of the data. Following the generation of these simulated scenarios, they can be analyzed to assess the probabilities of various outcomes, contingent upon variations in the involved parameters.

Subsequently, we will present a multivariate simulation employing the Monte Carlo technique. The parameters that will remain constant across the various simulated scenarios are as follows:

- the value of r, fixed at 0.3, which reasonably approximates a hypothetically plausible relationship between two variables;
- the number of simulations, referred to as iterations, conducted for each scenario, specifically set at $B = 1000$.

Conversely, the parameters of the simulation that will be varied to represent the different scenarios that researchers may realistically encounter when designing an experiment include:

- the values of d_1 and d_2, which will allow for an examination of how the effect of the mediator on the outcome fluctuates based on the intensity of the experimental effect on the mediator;

- the sample size n for each of the two groups, which will be maintained at an equal size.

The investigation will concentrate on analyzing these variations in terms of graphical visualizations of outcomes, statistical power and variability between the simulations.

4.2.3 Procedure for conducting the simulation in R

A preliminary step prior to conducting the simulation involved the creation of a new function in R, designated as "dataes". The corresponding R script is available in Appendix A. This function facilitates the generation of a list that presents a customizable number of data frames, encompassing two datasets: one for the experimental group, denoted as "T", and the other for the control group, referred to as "C". The two datasets consist of a series of multivariate normal distributions of random data, structured in accordance with the properties previously outlined. As a result, these distributions are centered on d_1 and d_2, with a covariance of r and a variance of 1. This approach enables the simulation of distributions for both the experimental and control groups with respect to the two effects of interest: the impact of exposure to the prosocial utility value condition on prosocial affordance beliefs and the subsequent direct influence of the latter on interest towards scientific topics.

Subsequently, this function was employed to generate a series of simulated scenarios. Specifically, we chose to simulate a total of 294 distinct scenarios, each resulting from the combination of various values assigned to the parameters considered in the simulation:

- The values of the differences, articulated in terms of Cohen's d, were selected to be identical for both effects, thereby reflecting the values most frequently observed in the psychological literature (Altoè, 2020). Specifically, the values chosen to represent both the effect on the mediator d_1 and on the dependent variable d_2 are $d_1 = d_2 = (0, 0.1, 0.2, 0.35, 0.5, 0.8, 1)$. Consequently, this results in seven distinct

levels for these two parameters.

- The values representing the various sample sizes for each group were selected to correspond to small, medium and large sample sizes, specifically $n = (10, 50, 100, 300, 500, 1000)$. This selection is informed by the research of Pastore and Altoè (2013), resulting in six distinct levels for this parameter.

Finally, it is important to note that the parameter r was fixed at 0.3, while the number of iterations for each scenario was set at $B = 1000$. Consequently, the combination of each level of the variable parameters and the specified number of repetitions for each scenario, which amount to $6 \times 7 \times 7 \times 1000$, yield a total of 294,000 datasets, which will be discussed in the subsequent section.

4.3 Simulating different scenarios

In this section, we describe the setup of the simulation. The first part focuses on examining the effect of manipulating the independent variable on the mediator variable, independent of its impact on the outcome variable. The second part addresses the effect of this manipulation on the dependent variable, relative to a specific, arbitrarily chosen value of the effect on the mediator.

For each of the two effects examined, we report an estimate of the effect size, a check for Type I error and an assessment of statistical power. Regarding the effect size, we used Hedges' g as an index to provide an unbiased estimate in relation to the δ established across simulations with different group sizes. In addition, confidence intervals (CIs) are presented to reflect the variability of the estimates across each group of simulations. By definition, confidence intervals express the uncertainty around the estimate of a population parameter. In this case, they represent the range of probable values for the parameter under study, i.e., the mean value of the effect size, g, for each simulation. Type I error is checked in scenarios where the true effect is assumed to be

zero, with the aim of calculating the probability at which H_0 is incorrectly rejected, using a nominal α value of 0.05. With respect to statistical power, we assess how it aligns with the commonly desirable threshold of 80% (Cohen, 1988).

For each of these three aspects, we present the simulation results both in tabular and graphical formats, as the visual representation provides an intuitive understanding of the key comparisons between different scenarios. Appendix A includes tables detailing the results obtained from the data simulations together with the R code.

4.3.1 Results of the effect on the mediator

4.3.1.1 Estimate and confidence interval of the effect on the mediator

Both the graph 4.2 and the table 6.1 clearly demonstrate that Hedges' g serves as an unbiased estimator of effect size, as the mean values of the estimates align closely with the Cohen's d values specified in the simulation. Additionally, it is evident that as the group size increases, the variability in the estimates decreases, as reflected by the narrowing of the confidence intervals. For the same n, the mean estimate does not seem to be influenced by the width of the interval.

An insightful observation from the graph is that even under the null hypothesis (H_0), where the true effect (set at 0 in this case) is null, there are instances where the estimated effect size appears notably high. For example, in cases where g is estimated at 0.8 with a sample size of 15. This underscores the importance of reporting not only the mean effect size but also the variability of the estimates, as the latter provides a fuller understanding of the potential range of effect sizes that could be observed.

4.3.1.2 Control of Type I error on the effect on the mediator

We condition the simulation under the null hypothesis (H_0) to examine the scenario in which the true effect, denoted as δ, is 0. The results, visible in figure 4.3, indicate that, irrespective of the sample size, the mean of the estimated values remains close

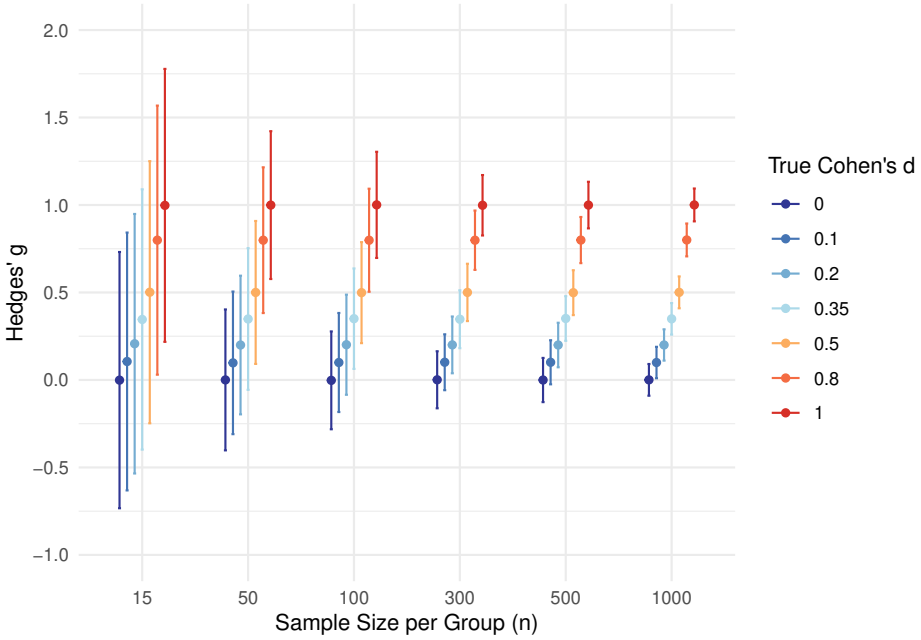

Figure 4.2: Mean estimated variable and variability of true Cohen's d through the unbiased effect size estimator Hedge's g for each of the different group sizes

to the nominal significance level of $\alpha = 0.05$. An important aspect to consider is that when the effect size is generated to equal 0, the interval exhibits substantial variability. Consequently, with such low sample sizes, there exists a risk of obtaining effect size estimates close to -1 or +1, which carries a clinical significance that is markedly different from that of observing an effect equal to 0. This phenomenon is considerably less prevalent when the sample size is increased.

4.3.1.3 Evaluation of power on the effect on the mediator

In this instance, we condition the analysis under the alternative hypothesis (H_1), examining the statistical power when the true effect size is different from zero. In figure 4.4 it can be observed that for small effect sizes, the statistical power is low, particularly when the sample size is also small. However, as the sample size increases, the power increases accordingly for each value of δ. This type of analysis allows researchers to determine, prior to data collection, how many participants are needed to achieve the

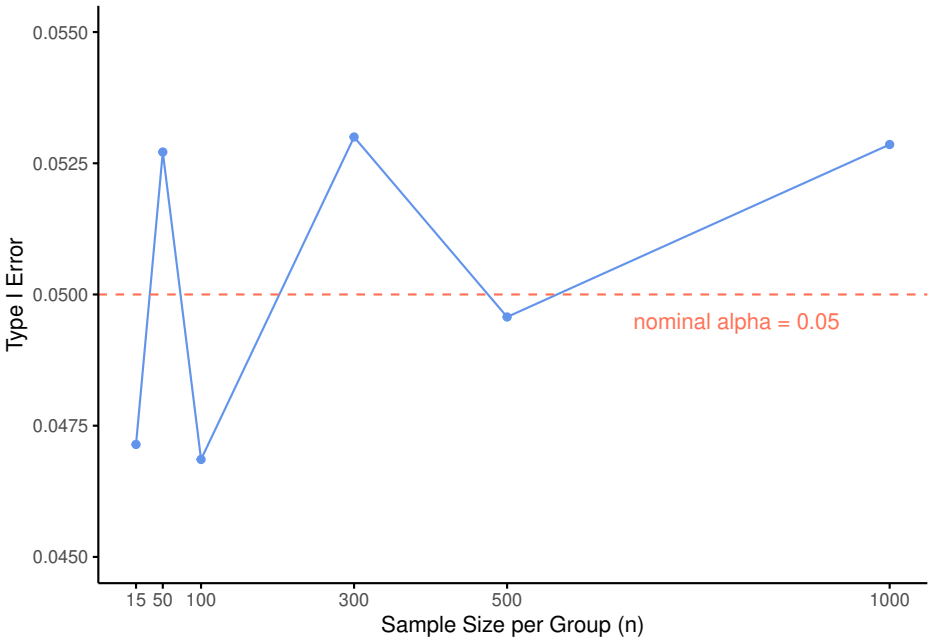

Figure 4.3: Control of Type I error on the effect on the mediator compared to the nominal significance level of 0.05

desired level of statistical power. It also helps in balancing the selection of the nominal value for Type I error (e.g., choosing $\alpha = 0.01$ instead of $\alpha = 0.05$ in cases where the power is already sufficiently high).

It is also noteworthy that, considering very small effects (e.g. $d = 0.1$), which nonetheless hold substantial relevance in psychology, even with exceptionally large samples, the maximum statistical power attained is only 60%. This observation should prompt researchers to critically reflect on the inferential risks associated with working with such small effect sizes. Furthermore, it should encourage them to transparently articulate the context in which they are conducting their research, thereby urging readers to interpret the results with due caution.

4.3.2 Results of the effect on the dependent variable

We now present the simulation results, focusing on the effect of the experimental manipulation on the dependent variable. For this analysis, we assume an effect size of

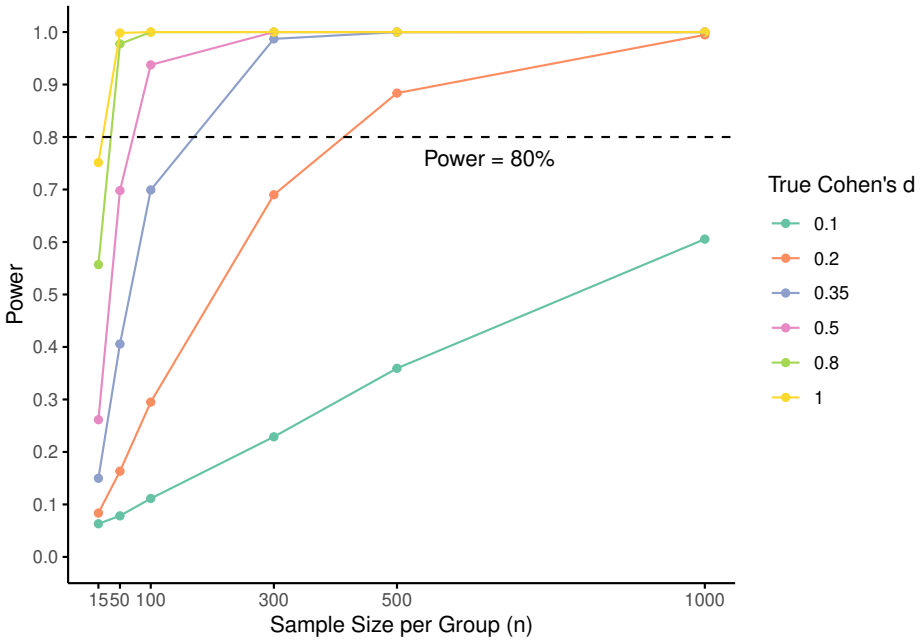

Figure 4.4: Evaluation of power on the effect on the mediator for each estimate of Cohen's d

$\delta = 0.35$ for the mediator, a value arbitrarily chosen as a reasonable average effect size for psychological studies when no other relevant information is available (Altoè et al., 2020).

The simulation outcomes mirror those observed in the analysis of the mediator effect. Specifically, the findings, visible in the following figures, remain consistent in terms of the mean effect size estimates, the variability of these estimates, Type I error control and statistical power. Upon analyzing the effect on the dependent variable, it becomes evident that small sample sizes are adequate for large effects. Conversely, in the case of small yet plausible effects in psychology, a maximum power of only 60% is achieved. Consequently, researchers should possess an awareness of the context in which they are operating and maintain transparency regarding this matter, thereby encouraging others to interpret their findings with appropriate caution.

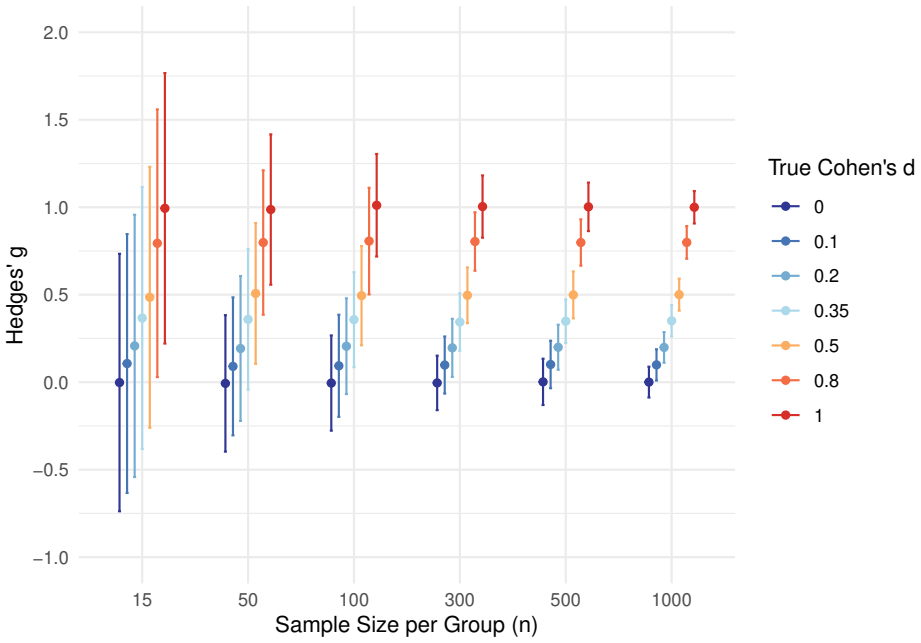

Figure 4.5: Mean estimated variable and variability of true Cohen's d on the dependent variable through the unbiased effect size estimator Hedge's g for each of the different group sizes. The narrowing of the confidence intervals reflects how, as the group size increases, the variability in the estimates decreases

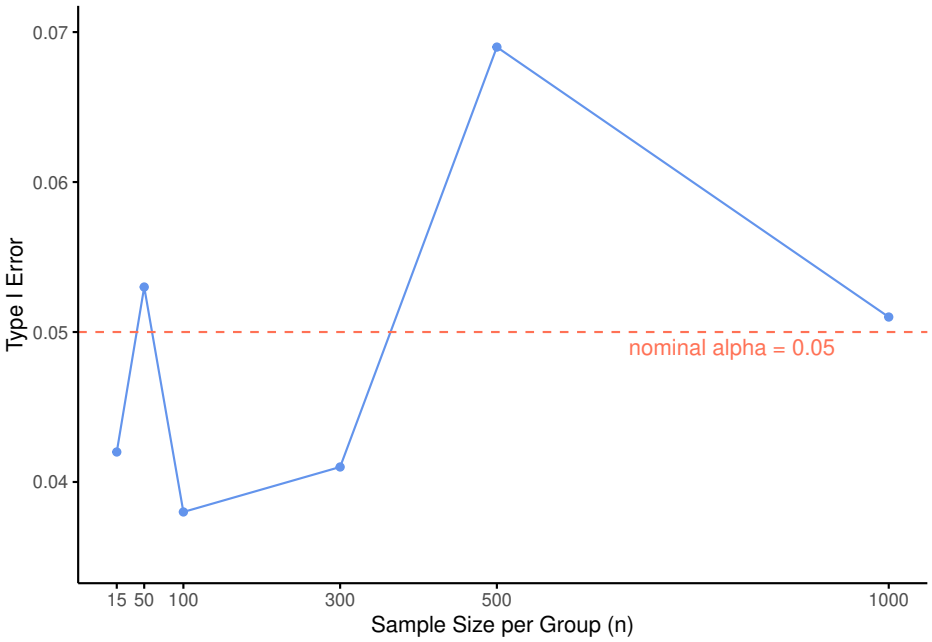

Figure 4.6: Mean estimated values of Type I error consistently distributed around the nominal significance level equal to 0.05, irrespective of the sample size

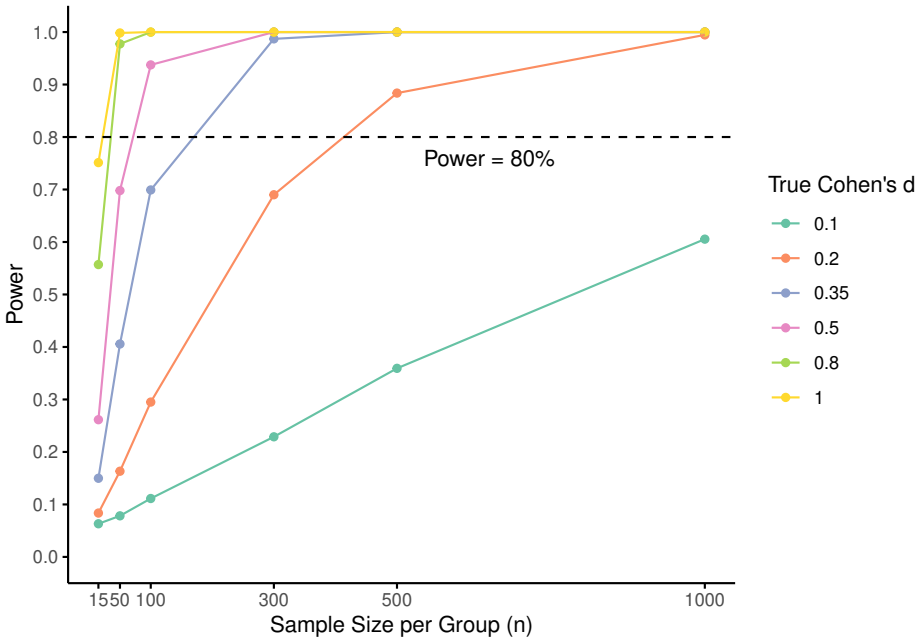

Figure 4.7: Evaluation of the power on the effect on the dependent variable for each estimate of Cohen's *d*. For small effect sizes, the statistical power is low, particularly when the sample size is also small. However, as the sample size increases, the power increases accordingly for each value of the true effect size

4.4 Applications of the simulation in research

As anticipated, in order to encourage the adoption of this approach, we offered the simulations reported in the preceding section for educational purposes. By illustrating various scenarios that researchers might encounter when conducting multivariate analyses, we demonstrated the functionality of the proposed framework using the full mediation model that represents the relationship between prosocial textbook content and prosocial affordance beliefs on science topic interest as a concrete example. Finally, it is essential to clarify how we intend for this framework to be practically implemented by researchers in their own work.

Among the key steps for conducting robust confirmatory research, as outlined in the first chapter, computer simulation plays a crucial role. A potential application of computer simulation to enhance rigorous research practices will be outlined subse-

quently. It is important to note, however, that simulations can be conducted at various stages within a study, serving distinct purposes. These additional applications will be discussed in greater detail in the following chapter.

Once the hypothesis has been properly formulated and translated into an appropriate mathematical model and after the variables of interest have been identified and the research design has been planned, the study design phase should conclude with a computer simulation. This step encourages researchers to critically consider the effect size under investigation. By reviewing the relevant literature and determining plausible effect sizes or setting the Smallest Effect Size of Interest (SESOI) (Anvari & Lakens, 2021), depending on the study's objectives, researchers can use these values to simulate potential outcomes. Through this process, researchers can visually observe how the effect size varies across simulations and calculate the study's statistical power. This, in turn, aids in planning the required sample size. Moreover, simulation allows for balancing the decision between Type I and Type II error rates, given their complementary roles in determining the validity of a study. By doing so, researchers can assess whether the study is adequately powered, ensuring its feasibility before actual data collection. Finally, once the data has been collected and the statistical analysis performed, researchers can compare the observed results with the predicted ones derived from the simulation. This comparison of both plots and values supports more meaningful interpretations and contributes to the falsification process, enhancing the study's overall validity.

We now intend to follow the same sequence of steps to demonstrate the initial implementation of the framework within the context of the study conducted by Zambrano et al. (2020). This study closely aligns with the simplified scenario we chose to present. Accordingly, the forthcoming chapter will be dedicated to the concrete application of the framework, representing the final step we believe is essential for advancing research practices - an objective that underpins the aims pursued throughout this dissertation.

Chapter 5

Case study

> *"Our contribution here is just [...] to remind experimenters that, once the data have been collected, the most important decisions have already been done."*
>
> — Gelman, 2023, p. 1

In this chapter, we outline the potential applications of simulation, distinguishing between simulations conducted prior to and those conducted following data collection and analysis, and detailing the procedures for implementing each approach. Specifically, as a concluding element of this dissertation, we aim to demonstrate the application of our proposed framework to a real hypothesis, illustrating the recommended steps for researchers to follow in advance of data collection. Following a detailed explanation of the simulation procedure, we offer our interpretation of the results, discuss their potential implications and outline the subsequent steps necessary to complete the study.

5.1 Different approaches to simulations

In the previous chapter, we examined a specific application of simulation within the context of confirmatory studies. Conducting simulation prior to data collection - thereby allowing researchers to effectively plan sample sizes and evaluate statistical power -

constitutes a prospective approach to simulation. Alternatively, simulation may be employed retrospectively, or a posteriori, following data collection. In this retrospective approach, effect sizes to estimate within the simulation are defined according to previous results in the literature or other information external to the study (Bertoldo et al., 2022), instead of being based on the observed results in the study, which is a widely-deprecated practice (Gelman, 2019).

In this thesis, we followed a prospective approach to simulation for educational purposes, aiming to promote a pre-testing mindset that strengthens confirmatory research, a practice we advocate. To illustrate this approach, the following section provides a comprehensive example of how to conduct a study using the proposed framework. Additionally, we provide a theoretical explanation of how this framework could be applied in studies adopting a retrospective approach to simulation.

5.2 Prospective approach to simulations: a case study

We intend to apply our framework to one of the hypotheses proposed by Zambrano et al. (2020), as our goal is to demonstrate its application within a simple multivariate context, consistent with the setting of their study. Specifically, we aim to use as a case study the phenomenon they examined: how the prosocial condition of a textbook predicts interest in scientific topics through prosocial affordance beliefs, thereby testing the complete mediation model that describes this relationship.

As researchers, after a thorough conceptualization of the phenomenon and the variables integral to the effect under study, we can proceed to configure the study and plan the framework's implementation. The initial step concerns conducting data simulation, followed by analyzing and discussing the anticipated results. Subsequently, we would undertake data collection and analysis, culminating in an interpretation of the findings by comparing them with the simulation outcomes, that is to say with the

results of our theoretical predictions.

5.2.1 Implementation of the framework

As discussed, all research must begin with a thoughtful consideration of the phenomenon of interest and the variables deemed relevant to its examination. Following this, it is essential to develop statistical models that align with the theoretical ones. In the case study that follows, we will build on previous discussions of the research object outlined in earlier chapters, with additional details available in Zambrano's original study (2020). Drawing on their multivariate design, the subsequent section provides a step-by-step outline of the actions required by our framework before and after data collection. In line with Gelman's recommendation to precede any data collection with computer simulation (2023), we will establish Cohen's d values that reflect plausible predictions of the true effect size for the phenomenon under study, thus providing a sound basis for the simulation. Identifying the true effect size, eventually designing a study that is both informative and falsifiable, can be approached in two main ways: by determining plausible effect sizes (Bertoldo et al., 2022) or by employing the smallest effect size of interest (SESOI) (Anvari & Lakens, 2021). A plausible effect size involves estimating the expected effect in the population, based on a comprehensive review of the literature (Bertoldo et al., 2022) and/or consultations with field experts, which should follow a rigorous protocol as detailed by O'Hagan (2019). Alternatively, rather than prioritizing plausibility, employing SESOI enables a focus on identifying a minimum effect size that holds clinical relevance (Anvari & Lakens, 2021).

Following the previously outlined approaches, we proceed with data simulation by setting effect sizes independently for both the mediator and the dependent variable, incorporating a single correlation hypothesis between the two. It is recommended in this context to consider multiple effect sizes as plausible predictions of the true effect, allowing for the evaluation of results across varying scenarios (Bertoldo et al., 2022).

Given that the manipulation's effect primarily impacts the mediator and that this effect subsequently transfers to the dependent variable - as observed in Zambrano et al. (2020) - we anticipate a stronger influence on the mediator, while expecting a comparatively smaller effect on the dependent variable. Thus, we selected medium to large Cohen's d values as plausible effect sizes to represent the effect on the mediator and small to medium Cohen's d values for the dependent variable's effect. Consequently, the simulation parameters that will vary to capture realistic scenarios include:

- the chosen values of d_1 (i.e. the effect on the mediating variable) and d_2 (i.e. the effect on the dependent variable), which will allow for an examination of how the effect of the mediator on the outcome fluctuates based on the intensity of the experimental effect on the mediator and, specifically, are $d_1 = (0.35, 0.5, 0.8)$ and $d_2 = (0.1, 0.2, 0.35)$;
- the sample size n for each of the two groups, which will be maintained at an equal size and represent the various sample sizes for each group in order to correspond to small, medium and large sample sizes, specifically $n = (50, 100, 150, 200, 250, 300, 350, 400)$.

Conversely, the parameters that will remain constant across the various simulated scenarios are as follows:

- the value of r, fixed at 0.3, which reasonably approximates a hypothetically plausible relationship between two variables;
- the number of simulations, referred to as iterations, conducted for each scenario, specifically set at $B = 1000$.

The investigation will concentrate on analyzing these variations in terms of graphical visualizations of outcomes, statistical power and variability between the simulations. Consequently, the combination of each level of the variable parameters and the specified number of repetitions for each scenario, which amount to $3 \times 3 \times 8 \times 1000$, yield a

simulation of 72,000 datasets, which will be discussed in the subsequent section.

5.2.2 Data simulation

At this stage, our objective is to examine the statistical power yielded by the simulation across various scenarios, each corresponding to a different plausible effect size. Based on the simulation results, we will be able to determine which study conditions would enable the detection of the expected effect size through data analysis. Following this, we will present the simulation results graphically and provide an interpretive discussion, with supplementary tables and the R code included in Appendix B.

Figure 5.1 represents the evaluation of the power on the effect on the mediator for each estimate of d_1. The variation in statistical power is illustrated according to selected group sizes. Specifically, it is evident that when n is low, statistical power tends to remain modest, particularly for smaller effect sizes. A reasonably high power level (> 0.80) is achieved when the group size reaches at least 150 participants, which is adequate for detecting medium effects (e.g., $d = 0.35$) as well as the more optimistic effect sizes.

At this stage, as researchers, we recognize that to ensure a sufficiently high power level, the initial simulation results suggest maintaining a minimum of 150 participants per group. Consequently, from this point forward, we will focus our evaluation and discussion on scenarios where $n > 150$.

Figure 5.2 represents the evaluation of the power on the effect on the dependent variable for each estimate of d_2. We observe that achieving a reasonable power level is challenging when the predicted effect size is small. Specifically, for a Cohen's d of 0.1, we infer that the effect size may be too small to reliably detect in our study. Conversely, for an effect size of $d = 0.2$, a substantial sample size (400 participants per group) is necessary to reach adequate power. For an optimistically anticipated effect size of $d = 0.35$, however, we can be reasonably confident in our ability to detect an effect on the

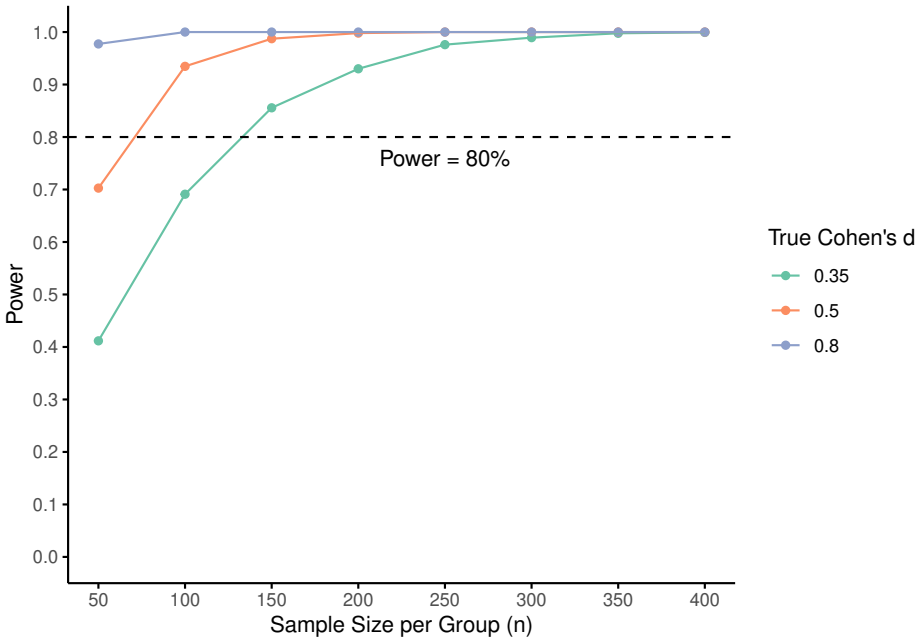

Figure 5.1: Evaluation of power on the effect on the mediator for each estimate of Cohen's d

dependent variable.

5.2.3 Discussion

At this juncture, as researchers preparing for the data collection phase, the simulation results indicate two viable approaches. The first option involves collecting data from 250 participants per group, which would yield a high probability of detecting each predicted effect on the mediator, substantial power for the more optimistic effect size or a moderate 60% power for detecting an effect size of 0.2 on the dependent variable. Alternatively, if resources permit, a second approach would involve collecting data from 400 participants per group, although even this sample size would not provide sufficient power to reliably detect an effect as a Cohen's d of 0.1.

Upon selecting the preferred approach - determining the number of participants to recruit - we would proceed with data collection and analysis. The final step would involve comparing the observed results with the outcomes hypothesized during the data

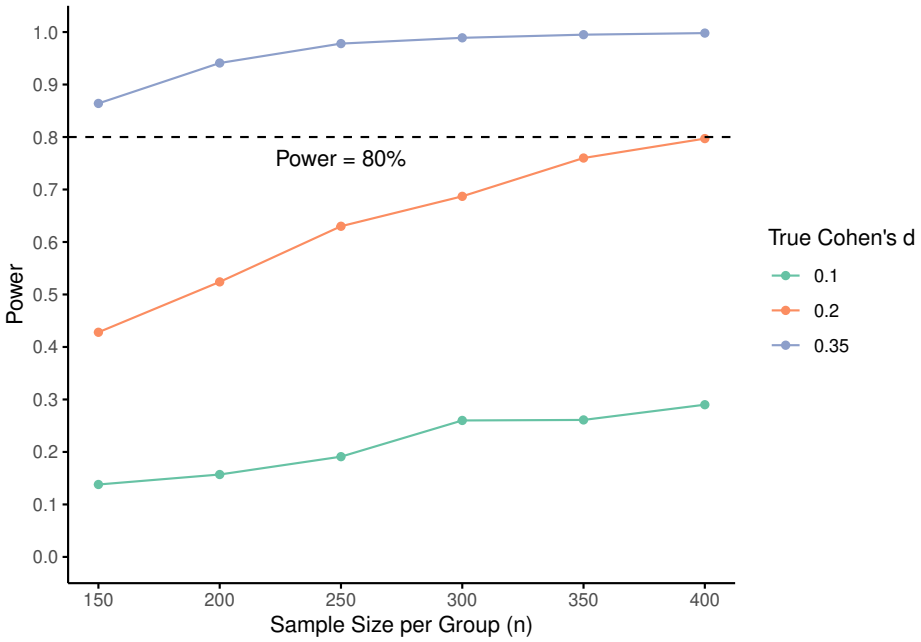

Figure 5.2: Evaluation of power on the effect on the dependent variable for each estimate of Cohen's d, considering exclusively $n > 150$

simulation phase and reflecting on any observed discrepancies.

5.3 Retrospective approach to simulations

In cases where data have already been collected and analyzed, a researcher may choose to conduct a simulation retrospectively, or a posteriori. It is essential in this context to base the simulation on plausible effect sizes or effect sizes of specific interest.

Once appropriate effect sizes are determined, the researcher can proceed with the simulation by following the procedures outlined in previous examples. This involves analyzing the estimated effect size in comparison with the plausible or targeted effect size and assessing the statistical power accordingly. Subsequently, the researcher may compare the findings from the initial data analysis - preferably displayed graphically - with the estimates derived from the simulation. In conclusion, the objective of evaluating the results of a completed study through retrospective simulations is not to simply

determine the significance of the observed effect. Rather, by grounding the retrospective analysis in plausible effect sizes, it aims to provide a more nuanced assessment of the results, focusing particularly on the observed effect size and, critically, the statistical power. In essence, this approach shifts the focus from the significance of the effect itself to an evaluation of the experimental design, thereby allowing for a more precise and informed interpretation of the findings.

5.4 Final considerations

The chapter presents the application of our framework to a concrete hypothesis, demonstrating the step-by-step approach we recommend for researchers. First, we emphasized that effect size must be considered from the outset of the study and positioned at the core of hypothesis formulation to ensure that predictions are truly falsifiable. Additionally, we aim for this study to contribute to a more nuanced understanding of effect size, as we illustrate how Cohen's d can serve to represent effect magnitudes in multivariate contexts - frequently applicable to hypotheses involving complex psychological phenomena. Furthermore, employing simulation prospectively, as previously shown, enables researchers to develop a realistic and testable understanding of the effect under investigation. This approach is further enhanced by graphical tools, which we recommend and have illustrated, to aid in the interpretation and visual representation of findings.

Overall, these results underscore the value of adopting this framework, supporting Gelman's assertion that "once the data have been collected, the most important decisions have already been made" (Gelman, 2023, p. 1). This framework thus offers an example of rigorous research practice that can yield more robust and insightful results in the scientific field.

Chapter 6

Conclusions

Psychology is currently facing a substantial credibility crisis and restoring trust in the field necessitates the adoption of rigorous research practices and transparency. This dissertation arises from this need, aiming to address critical issues that undermine psychological research and have contributed to the theory and replicability crisis, ultimately affecting the credibility of scientific findings in the discipline. Specifically, our focus is twofold: first, addressing the inadequacies in the quality of psychological theories and second, tackling issues related to the interpretation of effect sizes and the proliferation of underpowered studies. To address these challenges, we propose a comprehensive framework designed to guide researchers in conceptualizing and modeling effect sizes, as well as conducting data simulations prior to data collection. This framework is presented as a structured sequence of steps that researchers can implement throughout the research process, with the goal of improving both the quality and reliability of their studies.

While the proposed framework represents a substantial advancement, we recognize that it possesses certain limitations that warrant further consideration and refinement. First, as previously noted, we utilized a full mediation model as the illustrative scenario for the simulation, operating under the assumption that there is no direct influence of

the predictor on the dependent variable, with the effect being fully mediated by the mediator variable. This specific model was intentionally chosen for its simplicity and educational value. However, it does not account for the potential presence of a direct influence of the predictor, a common feature in many psychological phenomena. Furthermore, our discussion was constrained to a basic mediation model. Nonetheless, numerous models that aim to represent psychological phenomena must also account for variables that function as moderators, whether categorical (e.g., gender, socio-economic status) or quantitative (e.g., level of reward). Finally, for the sake of simplicity, it must be also recognised that our simulations consistently considered only a single correlation hypothesis. Addressing these complexities is essential for a more comprehensive application of the framework in psychological research.

Rather than being viewed as limitations, these points may be considered as future directions for the further development of the proposed framework. Since our primary goal was to introduce a novel framework, we opted for simplicity as the most effective explanatory approach. However, future work could extend the framework to partial mediation models, models with moderator variables and models involving more than three variables. Applying the framework in such complex scenarios, which more closely reflect the hypotheses in psychological research, would provide valuable insights, particularly in the field of multivariate analysis. Additionally, a promising direction for future development could involve enhancing the analysis of study properties. As suggested by Gelman & Carlin (2014), in addition to traditional Type I and Type II errors, two other errors - Type M (magnitude) and Type S (sign) - should be considered. Type M errors refer to the average overestimation of a statistically significant effect, while Type S errors represent the likelihood of obtaining a significant result in the opposite direction of the hypothesized effect. Incorporating the proposed framework within the context of “design analysis” (Gelman & Carlin, 2014), which broadly identifies the analysis of the properties of different studies, such as their statistical power as well as Type M

and Type S errors (Altoè et al., 2020), would offer a more comprehensive assessment of research outcomes.

Although, as Ioannidis (2005) argued, the ideal "gold standard" in research practices remains unattainable, we aspire for the proposed framework to serve as a modest example and inspiration for following good research practices. We therefore hope it could reinforce the importance of emphasizing effect size and statistical power, particularly within the field of psychology, and foster a broader awareness of their significance in producing more reliable and robust findings.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology*, *10*, 2893.

American Psychological Association. (2010). *Publication manual of the american psychological association*. American Psychological Association.

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2023). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*, *18*(2), 503–507.

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159.

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

Ballen, C. J., & Salehi, S. (2021). Mediation analysis in discipline-based education research using structural equation modeling: Beyond "what works" to understand how it works, and for whom. *Journal of Microbiology & Biology Education*, *22*(2),

10–1128.

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.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Bertani, A., Di Paola, G., Russo, E., & Tuzzolino, F. (2018). How to describe bivariate data. *Journal of Thoracic Disease*, *10*(2), 1133.

Bertoldo, G., Zandonella Callegher, C., Altoè, G., et al. (2022). Designing studies and evaluating research results: Type m and type s errors for pearson correlation coefficient. *Meta-Psychology*, *6*, 1–18.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Boring, E. G. (1919). Mathematical vs. Scientific significance. *Psychological Bulletin*, *16*(10), 335.

Coe, R. (2002). It's the effect size, stupid. *British Educational Research Association Annual Conference*, *12*, 14.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

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

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid cohen's "small,""medium," and 'large'for power analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207.

Cummins, R. (2000). *How does it work? Versus what are the laws?: Two conceptions of psychological explanation.*

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Bloomsbury Publishing.

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

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69–77.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Gall, T., Ioannidis, J. P., & Maniadis, Z. (2017). The credibility crisis in research: Can economics tools help? *PLoS Biology*, *15*(4), e2001846.

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9–e10.

Gelman, A. (2023). *Before data analysis: Additional recommendations for designing experiments to learn about the world.*

Gelman, A. (2024). Before data analysis: Additional recommendations for designing experiments to learn about the world. *Journal of Consumer Psychology*, *34*, 190–191.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1990). *The empire of chance: How probability changed science and everyday life* (Vol. 12). Cambridge University Press.

Glass, D. C., & Singer, J. E. (1972). *Urban stress: Experiments on noise and social stressors*. Academic Press.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications*. Routledge.

Hammond, S. (2000). Introduction to multivariate data analysis. *Research Methods in Psychology*, *2*, 272–396.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, *93*, 104082.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kimmelman, J., Mogil, J. S., & Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, *12*(5), e1001863.

Koeske, G. F. (1992). Moderator variables in social work research. *Journal of Social Service Research*, *16*(1-2), 159–178.

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648.

Lakens, D. (2022). *Improving your statistical inferences*.

Lakens, D. (2023). Methods-review boards could avert wasted research: Universities should ensure that study designs can actually answer their research questions. *Na-*

ture, *613*, 9.

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.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917.

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221080396.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*(4), 386.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. (1978). Theoretical risks and tabular astgris~ sir karl. *Sir Ronald, and the*.

Miola, A. C., & Miot, H. A. (2022). Comparing categorical variables in clinical and experimental studies. In *Jornal Vascular Brasileiro* (Vol. 21, p. e20210225). SciELO Brasil.

Neyman, J. (1957). " inductive behavior" as a basic concept of philosophy of science. *Revue De L'Institut International De Statistique*, 7–22.

Nosek, B. A., & Lakens, D. (2016). *Registered reports: A method to increase the credibility of published reports*.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, *73*(sup1), 69–81.

Otgaar, H., Riesthuis, P., Ramaekers, J. G., Garry, M., & Kloft, L. (2022). The

importance of the smallest effect size of interest in expert witness testimony on alcohol and memory. *Frontiers in Psychology*, *13*, 980533.

Pastore, M., & Altoè, G. (2013). Bayes factor e p-value: Così vicini, così lontani. *Giornale Italiano Di Psicologia*, *40*(1), 175–194.

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

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208.

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to götz et al.(2022). *Perspectives on Psychological Science*, *18*(2), 508–512.

Reinhold, S., Gegenfurtner, A., & Lewalter, D. (2018). Social support and motivation to transfer as predictors of training transfer: Testing full and partial mediation using meta-analytic structural equation modelling. *International Journal of Training and Development*, *22*(1), 1–14.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research*. Macmillan Publishing Co, Inc.

Tukey, J. W. (1987). We need both exploratory and confirmatory. *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1965-1986*, *4*, 811.

Ummul-Kiram, K., Silverstein, P., & Moin, S. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, *7*(1).

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860.

VandenBos, G. R. (Ed.). (2007). *APA dictionary of psychology*. American Psychological Association.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. van der, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*(253), 19–34.

Zambrano, J., Lee, G. A., Leal, C. C., & Thoman, D. B. (2020). Highlighting prosocial affordances of science in textbooks to promote science interest. *CBE—Life Sciences Education*, *19*(3), ar24.

Appendix A

The following R script was utilized to implement the proposed framework and to generate the graphical results discussed in Chapter 4.

```
# definition of a function to run the simulations
dataes<-function(n_t, n_c=n_t, dx, dy, r=.3, iter=10)
  {
require(MASS)
results <- list()
for (i in 1:iter){
#
data_t=mvrnorm(n=n_t, mu=c(dx,dy), Sigma=matrix(c(1,r,r,1),2,2) )
data_c=mvrnorm(n=n_c, mu=c(0,0), Sigma=matrix(c(1,r,r,1),2,2) )
#
d<-rbind(data_t,data_c)
group=factor( rep( c("T","C"), c(n_t,n_c)), levels = c("T", "C") )
d<-data.frame(d,group)
names(d)[1:2]<-c("x","y")
#
results <- c( results, list(d) )
                    }
```

```
return(invisible(results))
}


####################################
# implementation of the framework
# simulation setting
set.seed(23)
ng <- c(15, 50, 100, 300, 500, 1000)  # different group size
dm <- c(0, .1, .2, .35, .5, .8, 1)    # effect size on M
ddv <- dm                             # effect size on Y
B = 1000
results<-list()
#
for (i in ng)
  for (j in dm)
    for (k in ddv)
    {
    temp<-dataes(n_t=i,dx=j,dy=k,r=.3,iter = B)
    results <- c( results, list(temp) )
    }
d<-results
save(d,file="d.rda")


############################################
# formatting the results
index <- 1  # Indice per scorrere le combinazioni
```

```
for (i in ng) {
  for (j in dm) {
    for (k in ddv) {
      attr(d[[index]], "params") <- list(ng=i, dm=j, ddv=k)
      index <- index + 1
    }
  }
}
save(d,file="d.rda")
```

```
load("d.rda")


# defining useful functions
# function to compute Hedges' g
hedges_g <- function(m1, m2, sd1, sd2, n1, n2) {
  pooled_sd <- sqrt(((n1-1)*sd1^2 + (n2-1)*sd2^2) / (n1+n2-2))
  d <- (m1 - m2) / pooled_sd
  J <- 1 - (3 / (4*(n1 + n2 - 2) - 1))  # Hedges' correction
  g <- d * J
  return(g)
}
# function that calculates what we are interested in
calcola_quantita <- function(dataset) {
  # split the data by group
  data_t <- dataset[dataset$group == "T", ]
  data_c <- dataset[dataset$group == "C", ]
```

```r
  # group size
  n_t <- nrow(data_t)
  n_c <- nrow(data_c)

  # mean difference
  mean_diff_x <- mean(data_t$x) - mean(data_c$x)
  mean_diff_y <- mean(data_t$y) - mean(data_c$y)

  # t-test
  t_test_x <- t.test(x ~ group, data=dataset)$p.value
  t_test_y <- t.test(y ~ group, data=dataset)$p.value

  # Hedges' g on X e Y
  g_x <- hedges_g(mean(data_t$x), mean(data_c$x),
                  sd(data_t$x), sd(data_c$x), n_t, n_c)
  g_y <- hedges_g(mean(data_t$y), mean(data_c$y),
                  sd(data_t$y), sd(data_c$y), n_t, n_c)

  return(c(t_test_x, t_test_y, g_x, g_y))
}
############################
############################
# initialising an empty data.frame for results
results_df <- data.frame()

# loop over each combination of results in 'd'
for (idx in 1:length(d)) {
```

```
  simulazioni <- d[[idx]]
# extracting the simulation group for the current combination

  # initialising vectors to store the results
  p_value_x_vec <- numeric(length(simulazioni))
  p_value_y_vec <- numeric(length(simulazioni))
  g_x_vec <- numeric(length(simulazioni))
  g_y_vec <- numeric(length(simulazioni))

  # computing the quantities of interest for each simulation
  for (b in 1:length(simulazioni)) {
    quantita <- calcola_quantita(simulazioni[[b]])

    # assigning results to respective vectors
    p_value_x_vec[b] <- quantita[1]
    p_value_y_vec[b] <- quantita[2]
    g_x_vec[b] <- quantita[3]
    g_y_vec[b] <- quantita[4]
  }

  # calculating the power for the t-test (p-value proportion < 0.05)
  potenza_x <- mean(p_value_x_vec < 0.05)
  potenza_y <- mean(p_value_y_vec < 0.05)

  # mean Hedges' g and its variance
  hedges_g_medio_x <- mean(g_x_vec)
```

```
  hedges_g_medio_y <- mean(g_y_vec)
  var_g_x <- var(g_x_vec)
  var_g_y <- var(g_y_vec)

  comb_params <- attr(d[[idx]], "params")

  # creating a row of results
  new_row <- data.frame(ng=comb_params$ng,
                        dm=comb_params$dm,
                        ddv=comb_params$ddv,
                        potenza_x=potenza_x,
                        potenza_y=potenza_y,
                        hedges_g_medio_x=hedges_g_medio_x,
                        hedges_g_medio_y=hedges_g_medio_y,
                        var_g_x=var_g_x, var_g_y=var_g_y)

  # adding a row to the final data frame
  results_df <- rbind(results_df, new_row)
}

ris<-results_df
save(ris,file="ris.rda")
```

The tables below present the results from applying the proposed framework, as depicted graphically in Chapter 4. These results pertain to simulations based on true effect sizes predicted in terms of Cohen's d across various group sizes.

Table 6.1: Mean estimated variable and variability of true Cohen's d on the mediator through the unbiased effect size estimator Hedge's g for each of the different group sizes

True Cohen's d	Sample Size per Group (n)	Hedges' g	CI (lower bound)	CI (upper bound)	CI (width)
0.00	15	-0.001	-0.733	0.731	1.464
	50	0.000	-0.402	0.403	0.805
	100	-0.002	-0.282	0.277	0.559
	300	0.001	-0.162	0.164	0.326
	500	0.000	-0.127	0.126	0.253
	1000	0.001	-0.090	0.091	0.181
0.10	15	0.106	-0.631	0.842	1.474
	50	0.098	-0.310	0.505	0.815
	100	0.100	-0.183	0.383	0.566
	300	0.101	-0.058	0.261	0.319
	500	0.101	-0.025	0.227	0.252
	1000	0.100	0.010	0.189	0.179
0.20	15	0.207	-0.534	0.949	1.483
	50	0.200	-0.197	0.596	0.792
	100	0.201	-0.085	0.487	0.572
	300	0.200	0.038	0.363	0.324
	500	0.200	0.073	0.327	0.254
	1000	0.200	0.111	0.289	0.179
	15	0.346	-0.398	1.090	1.488

	50	0.349	-0.057	0.754	0.810
	100	0.350	0.063	0.637	0.574
0.35	300	0.347	0.182	0.513	0.331
	500	0.351	0.224	0.479	0.255
	1000	0.349	0.259	0.439	0.180
	15	0.502	-0.248	1.251	1.498
	50	0.500	0.092	0.909	0.817
	100	0.499	0.211	0.788	0.577
0.50	300	0.500	0.337	0.664	0.328
	500	0.499	0.371	0.627	0.256
	1000	0.501	0.410	0.592	0.182
	15	0.799	0.030	1.568	1.538
	50	0.799	0.382	1.215	0.833
	100	0.799	0.504	1.093	0.589
0.80	300	0.799	0.629	0.968	0.339
	500	0.800	0.668	0.931	0.264
	1000	0.800	0.706	0.894	0.188
	15	0.998	0.218	1.778	1.560
	50	0.999	0.577	1.422	0.845
	100	1.001	0.697	1.304	0.607
1.00	300	0.999	0.826	1.171	0.345
	500	1.000	0.866	1.133	0.266
	1000	1.000	0.907	1.094	0.188

Table 6.2: Control of Type I error on the effect on the mediator compared to the nominal significance level of 0.05

Sample Size per Group (n)	**Type I Error**
15	0.047
50	0.053
100	0.047
300	0.053
500	0.050
1000	0.053

Table 6.3: Evaluation of the power on the effect on the mediator for each estimate of Cohen's d

True Cohen's d	**Sample Size per Group (n)**	**Power**
0.10	15	0.063
	50	0.078
	100	0.111
	300	0.229
	500	0.359
	1000	0.605
	15	0.084
	50	0.163
	100	0.295
	300	0.690
	500	0.884

0.20	1000	0.995
0.35	15	0.150
	50	0.406
	100	0.699
	300	0.987
	500	1.000
	1000	1.000
0.50	15	0.261
	50	0.698
	100	0.937
	300	1.000
	500	1.000
	1000	1.000
0.80	15	0.557
	50	0.977
	100	1.000
	300	1.000
	500	1.000
	1000	1.000
1.00	15	0.751
	50	0.998
	100	1.000
	300	1.000
	500	1.000
	1000	1.000

Table 6.4: Mean estimated variable and variability of true Cohen's d on the dependent variable through the unbiased effect size estimator Hedge's g for each of the different group sizes

True Cohen's d	**Sample Size per Group (n)**	**Hedges' g**	**CI (lower bound)**	**CI (upper bound)**	**CI (width)**
0.00	-0.002	15	-0.738	0.734	1.472
	-0.007	50	-0.397	0.383	0.780
	-0.005	100	-0.277	0.267	0.544
	-0.004	300	-0.160	0.152	0.312
	0.002	500	-0.130	0.134	0.265
	0.001	1000	-0.087	0.089	0.176
0.10	0.107	15	-0.633	0.847	1.480
	0.090	50	-0.304	0.485	0.789
	0.094	100	-0.198	0.386	0.584
	0.098	300	-0.065	0.261	0.326
	0.101	500	-0.034	0.237	0.271
	0.099	1000	0.010	0.188	0.178
0.20	0.207	15	-0.542	0.957	1.499
	0.193	50	-0.221	0.606	0.828
	0.206	100	-0.068	0.479	0.547
	0.196	300	0.030	0.362	0.332
	0.200	500	0.071	0.328	0.257
	0.199	1000	0.111	0.286	0.175
	0.367	15	-0.382	1.116	1.498

	0.359	50	-0.043	0.761	0.804
	0.358	100	0.086	0.629	0.543
0.35	0.344	300	0.180	0.508	0.329
	0.349	500	0.224	0.473	0.250
	0.351	1000	0.261	0.440	0.179
	0.485	15	-0.260	1.231	1.492
	0.507	50	0.105	0.910	0.805
	0.495	100	0.211	0.778	0.567
0.50	0.497	300	0.337	0.656	0.318
	0.499	500	0.365	0.633	0.268
	0.501	1000	0.409	0.592	0.183
	0.794	15	0.029	1.559	1.530
	0.798	50	0.386	1.211	0.825
	0.807	100	0.502	1.111	0.609
0.80	0.804	300	0.637	0.971	0.334
	0.798	500	0.666	0.931	0.265
	0.799	1000	0.706	0.892	0.186
	0.994	15	0.221	1.767	1.546
	0.987	50	0.557	1.417	0.859
	1.011	100	0.718	1.304	0.586
1.00	1.004	300	0.826	1.182	0.356
	1.002	500	0.864	1.141	0.277
	1.000	1000	0.907	1.092	0.185

Table 6.5: Control of Type I error on the effect on the dependent variable

Sample Size per Group (n)	**Type I Error**
15	0.042
50	0.053
100	0.038
300	0.041
500	0.069
1000	0.051

Table 6.6: Evaluation of the power on the effect on the depdent variable for each estimate of Cohen's *d*

True Cohen's d	**Sample Size per Group (n)**	**Power**
0.10	15	0.058
0.20	15	0.093
0.35	15	0.156
0.50	15	0.250
0.80	15	0.562
1.00	15	0.742
0.10	50	0.063
0.20	50	0.164
0.35	50	0.421
0.50	50	0.723
0.80	50	0.975
1.00	50	0.999

0.10	100	0.117
0.20	100	0.293
0.35	100	0.730
0.50	100	0.937
0.80	100	1.000
1.00	100	1.000
0.10	300	0.212
0.20	300	0.678
0.35	300	0.987
0.50	300	1.000
0.80	300	1.000
1.00	300	1.000
0.10	500	0.360
0.20	500	0.887
0.35	500	1.000
0.50	500	1.000
0.80	500	1.000
1.00	500	1.000
0.10	1000	0.608
0.20	1000	0.995
0.35	1000	1.000
0.50	1000	1.000
0.80	1000	1.000
1.00	1000	1.000

Appendix B

The following R script was utilized to implement the proposed framework and to generate the graphical results discussed in Chapter 5.

```
# definition of a function to run the simulations
dataes<-function(n_t, n_c=n_t, dx, dy, r=.3, iter=10)
  {
require(MASS)
results <- list()
for (i in 1:iter){
#
data_t=mvrnorm(n=n_t, mu=c(dx,dy), Sigma=matrix(c(1,r,r,1),2,2) )
data_c=mvrnorm(n=n_c, mu=c(0,0), Sigma=matrix(c(1,r,r,1),2,2) )
#
d<-rbind(data_t,data_c)
group=factor( rep( c("T","C"), c(n_t,n_c)), levels = c("T", "C") )
d<-data.frame(d,group)
names(d)[1:2]<-c("x","y")
#
results <- c( results, list(d) )
                    }
```

```
return(invisible(results))
}
####################
# implementation of the framework
# simulation setting
set.seed(23)
ng <- c(50, 100, 150, 200, 250, 300, 350, 400)  # different group size
dm <- c(.35, .5, .8)    # effect size on M
ddv <- c(.1,.2, .35)    # effect size on Y
8*3*3 # 72 combinations
B <- 1000 # 72000 dataset


results<-list()
#
for (i in ng)
  for (j in dm)
    for (k in ddv)
    {
    temp<-dataes(n_t=i,dx=j,dy=k,iter = B)
    results <- c( results, list(temp) )
    }
d<-results
# formatting the results
index <- 1
for (i in ng) {
  for (j in dm) {
```

```
    for (k in ddv){
      attr(d[[index]], "params") <- list(ng=i, dm=j, ddv=k)
      index <- index + 1
    }
  }
}
# check
attr(d[[2]], "params")
save(d,file="d.rda")
```

```
load("d.rda")

# defining useful functions
# function to compute Hedges' g
hedges_g <- function(m1, m2, sd1, sd2, n1, n2) {
  pooled_sd <- sqrt(((n1-1)*sd1^2 + (n2-1)*sd2^2) / (n1+n2-2))
  d <- (m1 - m2) / pooled_sd
  J <- 1 - (3 / (4*(n1 + n2 - 2) - 1))  # Hedges' correction
  g <- d * J
  return(g)
}
# function that calculates what we are interested in
calcola_quantita <- function(dataset) {
  # split the data by group
  data_t <- dataset[dataset$group == "T", ]
  data_c <- dataset[dataset$group == "C", ]
```

```r
  # groups size
  n_t <- nrow(data_t)
  n_c <- nrow(data_c)

  # mean difference
  mean_diff_x <- mean(data_t$x) - mean(data_c$x)
  mean_diff_y <- mean(data_t$y) - mean(data_c$y)

  # t-test
  t_test_x <- t.test(x ~ group, data=dataset)$p.value
  t_test_y <- t.test(y ~ group, data=dataset)$p.value

  # Hedges' g on X e Y
  g_x <- hedges_g(mean(data_t$x), mean(data_c$x),
                  sd(data_t$x), sd(data_c$x), n_t, n_c)
  g_y <- hedges_g(mean(data_t$y), mean(data_c$y),
                  sd(data_t$y), sd(data_c$y), n_t, n_c)

  return(c(t_test_x, t_test_y, g_x, g_y))
}
############################
############################
# initialising an empty data.frame for results
results_df <- data.frame()

# loop over each combination of results in 'd'
for (idx in 1:length(d)) {
```

```r
simulazioni <- d[[idx]]
# extracting the simulation group for the current combination

# initialising vectors to store the results
p_value_x_vec <- numeric(length(simulazioni))
p_value_y_vec <- numeric(length(simulazioni))
g_x_vec <- numeric(length(simulazioni))
g_y_vec <- numeric(length(simulazioni))

# computing the quantities of interest for each simulation
for (b in 1:length(simulazioni)) {
  quantita <- calcola_quantita(simulazioni[[b]])

  # assigning results to respective vectors
  p_value_x_vec[b] <- quantita[1]
  p_value_y_vec[b] <- quantita[2]
  g_x_vec[b] <- quantita[3]
  g_y_vec[b] <- quantita[4]
}

# calculating the power for the t-test (p-value proportion < 0.05)
potenza_x <- mean(p_value_x_vec < 0.05)
potenza_y <- mean(p_value_y_vec < 0.05)

# mean Hedges' g and its variance
hedges_g_medio_x <- mean(g_x_vec)
```

```
  hedges_g_medio_y <- mean(g_y_vec)
  var_g_x <- var(g_x_vec)
  var_g_y <- var(g_y_vec)

  comb_params <- attr(d[[idx]], "params")

  # creating a row of results
  new_row <- data.frame(ng=comb_params$ng, dm=comb_params$dm,
                        ddv=comb_params$ddv,
                        potenza_x=potenza_x,
                        potenza_y=potenza_y,
                        hedges_g_medio_x=hedges_g_medio_x,
                        hedges_g_medio_y=hedges_g_medio_y,
                        var_g_x=var_g_x, var_g_y=var_g_y)

  # adding a row to the final data frame
  results_df <- rbind(results_df, new_row)
}

ris<-results_df
save(ris,file="ris.rda")
```

Presented below are tables containing the results from applying the proposed framework, as illustrated graphically in Chapter 5. These results pertain to the evaluation of statistical power achievable based on the predicted effect sizes and varying group sizes.

Table 6.7: Evaluation of the power on the effect on the mediator for each estimate of Cohen's d

True Cohen's d	**Sample Size per Group (n)**	**Power**
0.35	50	0.412
	100	0.691
	150	0.856
	200	0.930
	250	0.976
	300	0.989
	350	0.998
	400	0.999
0.50	50	0.703
	100	0.935
	150	0.987
	200	0.998
	250	1.000
	300	1.000
	350	1.000
	400	1.000
	50	0.977
	100	1.000
	150	1.000
	200	1.000
	250	1.000
	300	1.000

0.80	350	1.000
	400	1.000

Table 6.8: Evaluation of the power on the effect on the dependent variable for each estimate of Cohen's *d*

True Cohen's d	**Sample Size per Group (n)**	**Power**
0.10	150	0.138
	200	0.157
	250	0.191
	300	0.260
	350	0.261
	400	0.290
0.20	150	0.428
	200	0.524
	250	0.630
	300	0.687
	350	0.760
	400	0.797
0.35	150	0.864
	200	0.941
	250	0.978
	300	0.989
	350	0.995
	400	0.998