UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Developmental Psychology and Socialisation

Ph.D. COURSE IN PSYCHOLOGICAL SCIENCES

XXXIV CYCLE

Towards the Formalization of Research Hypotheses in Psychology: Design Analysis and Model Comparison

Coordinator: Prof. Giovanni Galfano

Supervisor: Prof. Gianmarco Altoè

Co-Supervisor: Prof. Livio Finos

Ph.D. student: Claudio Zandonella Callegher

Abstract

The evaluation of research and theoretical hypotheses is one of the principal goals of empirical research. In fact, when conducting a study, researchers usually have expectations based on hypotheses or theoretical perspectives they want to evaluate according to the observed data. To do that, different statistical approaches have been developed, for example, the Null Hypothesis Significance Testing (NHST).

In psychology, the NHST is the dominant statistical approach to evaluate research hypotheses. In reality, however, the NHST approach does not allow researchers to answer the question they usually are interested in. In fact, the NHST approach does not quantify the evidence in favour of a hypothesis, but it only quantifies the evidence against the null hypothesis. This can easily lead to the misinterpretation of the results that, together with a mindless and mechanical application of the NHST approach, is considered as one of the causes of the ongoing replicability crisis.

In the first part of the thesis, we introduce the *Design Analysis* framework that allows us to evaluate the inferential risks related to effect size estimation when selecting for significance. In the case of underpowered studies evaluating complex multivariate phenomena with noisy data (all very common conditions in psychology), selecting for significance can easily lead to misleading and unreliable results. This aspect is often neglected in traditional power Analysis. Design analysis, instead, highlights this relevant issue.

In the second part of the thesis, we move away from the NHST towards the model comparison approach. Model comparison allows us to properly evaluate the relative evidence in favour of one hypothesis according to the data. First, research hypotheses are formalized into different statistical models, subsequently, these are evaluated according to different possible criteria. We consider the information criteria and the Bayes Factor with encompassing prior. Information criteria assess models predictive ability penalizing for model complexity. Bayes Factor with encompassing prior, instead, allows researchers to easily evaluate informative hypotheses with equality and inequality constraints on the model parameters.

Abstract (Italian)

La valutazione di ipotesi definite in accordo con le aspettative dei ricercatori o di prospettive teoriche è uno degli obiettivi principali della ricerca empirica. Quando viene condotto uno studio, infatti, i ricercatori di solito vogliono valutare la plausibilità delle loro ipotesi sulla base dei dati osservati. Per fare ciò, sono stati sviluppati diversi approcci statistici come, ad esempio, il Null Hypothesis Significance Testing (NHST).

In psicologia, il NHST è l'approccio statistico dominante per valutare le ipotesi di ricerca. In realtà, tuttavia, l'approccio NHST non consente ai ricercatori di rispondere alla domanda a cui di solito sono interessati. Infatti, l'approccio NHST non quantifica l'evidenza a favore di un'ipotesi, ma quantifica solo l'evidenza contro l'ipotesi nulla. Ciò può facilmente portare a un'errata interpretazione dei risultati che, insieme all'applicazione meccanica ad insensata dell'approccio NHST, è considerata una delle cause dell'attuale crisi di replicabilità.

Nella prima parte della tesi, introduciamo il framework della *Design Analysis* che ci permette di valutare i rischi inferenziali legati alla stima della dimensione dell'effetto quando si seleziona per la significatività. Nel caso di studi con campioni ridotti che valutano fenomeni complessi e con grande variabilità nei dati (tutte condizioni molto comuni in psicologia), la selezione per significatività può facilmente portare a risultati fuorvianti ed inaffidabili. Questo aspetto è spesso trascurato nella *Power Analysis* tradizionale. La Design Analysis, invece, mette in evidenza questo importante problema.

Nella seconda parte della tesi, ci spostiamo dal NHST verso l'approccio del *Model Comparison*. Il Model Comparison ci consente di valutare correttamente l'evidenza relativa a favore di un'ipotesi in base ai dati. In primo luogo, le ipotesi di ricerca vengono formalizzate sotto forma di diversi modelli statistici. Successivamente, queste vengono valutate secondo diversi possibili criteri come, ad esempio, gli *Information Criteria* e il *Bayes Factor con encompassing prior*. Gli Information Criteria valutano la capacità predittiva dei modelli penalizzando per la complessità del modello. Il Bayes Factor con encompassing prior, invece, consente ai ricercatori di valutare facilmente ipotesi informative con vincoli di uguaglianza e disuguaglianza sui parametri del modello.

A Luca e Sara

perché i fratelli ti accompagnano sempre

Un dottorato, 3 anni in cui tutto è cambiato.

Ringrazio...

Il Prof. Altoè che il dolore più grande ha affrontato

I miei genitori che tanto hanno sopportato

Sara e Daniel che le montagne hanno ritrovato

Luca e Denis che un nuovo percorso hanno disegnato

Baldo che alla mia stessa sorte è legato

Unz che di pannolini si è circondato

Novak che della sua passione è finalmente ripagato

Jake che un nuovo equilibrio ha trovato

Inoltre, ringrazio tutti i colleghi di Psicostat per lo stimolante ambiente che nel dipartimento hanno creato. In particolare, i Prof. Altoè (di nuovo, e lo dovrei ringraziare per altri mille motivi ancora), Finos, Pastore, Calcagnì, Girardi che della statistica mi hanno fatto innamorare. Ringrazio, Tatiana che del dottorato mi aveva avvisato. Ringrazio la Prof.ssa Agnoli, la Prof.ssa Benavides, Filippo, Giulia, Enrico ed il gruppo di lavoro della Prof.ssa Farroni per le bellissime collaborazioni.

Infine, ringrazio tutti i dottorandi, compagni di sventura, che hanno di certo reso migliori questi 3 anni. Ringrazio i coinquilini che hanno condiviso una parte di percorso e che porterò con me. Ringrazio i compagni di corda Parin e Jacopone perché la ricerca è la ricerca ma la montagna è tutt'altra cosa.

Contents

1 Introduction

> *All models are wrong but some are useful*
>
> — **Box (1979)**

If I ever get a tattoo, this quote would be the first choice, probably together with "`on CRAN`". Fortunately, I am afraid of needles so this will never happen and I will not have to care about having embarrassing tattoos. Nevertheless, Box's famous words are truly inspiring as they summarize the aim of applied statistics and scientific research more in general: we build models that approximate the reality, these are never perfect but they allow us to understand and predict what surround us.

In Psychology, however, sometimes it seems that we forgot about it. Thus, the purpose of this thesis is to reason about how we apply statistical inference to answer our research questions and the take-home message is simply "*stop testing start modeling*". Of course, this is provocative as there will never be a rule that applies to all cases but different situations would always require different solutions[1]. In this way, however, we hope to enhance researchers' awareness about the issues related to the misuse of statistical techniques: statistical inference should not be used mechanically as a black box but it is important to understand its mechanisms and properly applying the different available statistical techniques. We think that

[1] This is not intended to be a rule otherwise we enter the *Barber Paradox* ;). See https://en.wikipedia.org/wiki/Barber_paradox

this becomes more natural when thinking about modeling the data rather than testing it, but, to clarify this let's start from the beginning.

1.1 A Three Years Long Journey into Statistical Inference

Considering a thesis uniquely as a sum of studies is extremely reductive, there is much more going on underneath. A thesis is the result of a three years-long process in which a PhD student grows up becoming a young researcher. This is an extremely stimulating period in which the formation received and the surrounding academic culture will direct future researchers' attitudes. Thus, to properly understand the meaning and reasons behind a thesis, it is important to consider the circumstances during its creation.

Before starting my PhD, I thought these were really exciting times for research in psychology. During the university courses, professors presented us with many interesting studies about brain functioning and any type of psychological or social behaviours. I thought we were close to revealing many mysteries of the human mind, so what better moment to start a PhD than this? Well, as soon as I started my PhD, I find out a pretty uncomfortable truth: we were in the middle of the replicability crisis.

1.1.1 Replicability Crisis

Following Ioannidis (2005) "*Why Most Published Research Findings Are False*" paper, there has been a growing concern about the reliability of the results in psychology and social sciences. Many studies failed to replicate previous results in the literature or, if replicated, the estimated effects were much smaller than the original ones (Camerer et al., 2018; Open Science Collaboration, 2015). To refer to this issue, researchers started using the term "*replicability crisis*", a phenomenon that is not limited to psychological and social sciences but, as suggested by different surveys, involves many other scientific fields (Baker, 2016).

This ongoing replicability crisis created a huge debate in the literature regarding Questionable Research Practices (QRPs; John et al., 2012), questionable measurement practices (Flake & Fried, 2020; Schimmack, 2021), and other methodological issues (Pashler & Wagenmakers, 2012; Stangor & Lemay, 2016). In particular, the use of Null Hypothesis Significance Testing (NHST) procedure has been strongly criticized (Cumming, 2014; Greenland et al., 2016; Nuzzo, 2014; Wasserstein et al., 2019).

1.1.2 Null Hypothesis Significance Testing

The Null Hypothesis Significance Testing (NHST) is the dominant statistical approach in psychological and social research (Chavalarias et al., 2016). Surprisingly, however, the NHST does not properly exist in statistical sciences but it is given by the combination of two different conflicting approaches: Fisher's significance testing and Neyman and Pearson's hypothesis testing (Gigerenzer et al., 2004; Perezgonzalez, 2015).

Without going into details, the resulting approach has several limitations, these will be discussed in the next chapters and interested readers can refer to Greenland et al. (2016) and Szucs and Ioannidis (2017). In particular, NHST mindless application, in what Gigerenzer et al. (2004) defined as the " *Null Ritual*", can often lead to misleading and unreliable results. For example, imagine that we would like to evaluate the efficacy of a new psychological treatment. To do that we compare a treatment group and a control group, each one formed by 20 participants. Suppose we observe a Cohen's d of .8. In this case, in a two-tailed t-test with a traditional α value of .5, we would obtain a statistical significant result ($t(38) = 2.53, p = .016$), see Figure 1.1.

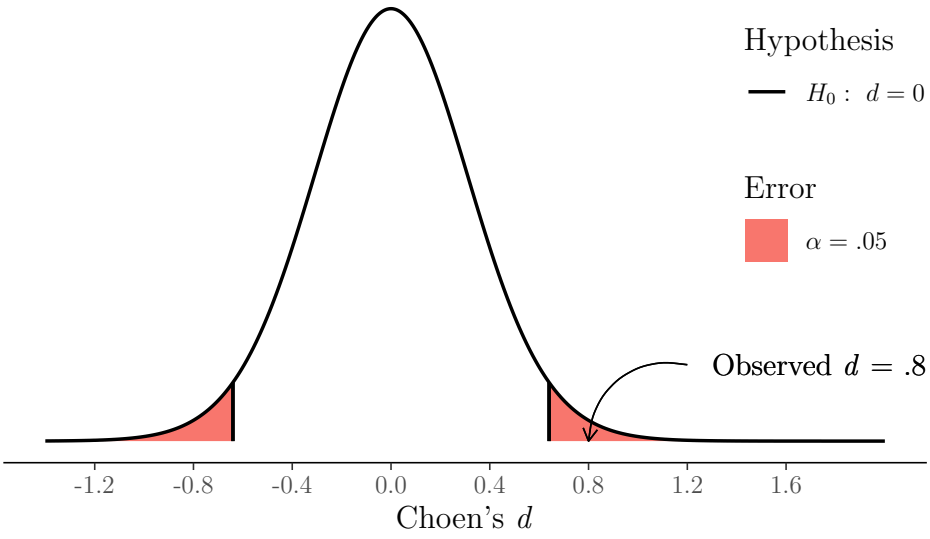

Figure 1.1: NHST comparing a treatment group and a control group ($n_{tot} = 40$), distribution under $H_0: \ d = 0$.

At this point, we would be really happy as we have found a statistically significant result and we can publish it. However, we have forgotten to consider another important element of statistical inference, which is statistical power. The NHST does not require the formalization of the alternative hypothesis H_1 but only the

null-hypothesis H_0 is needed to compute the p-value. For this reason, the alternative hypothesis is often neglected by the researchers until they have to compute the power. Suppose that, after an extensive literature review, we consider as a plausible effect size $H_1: \ d = .25$ and we find out that the resulting power is only 12%, see Figure 1.2.

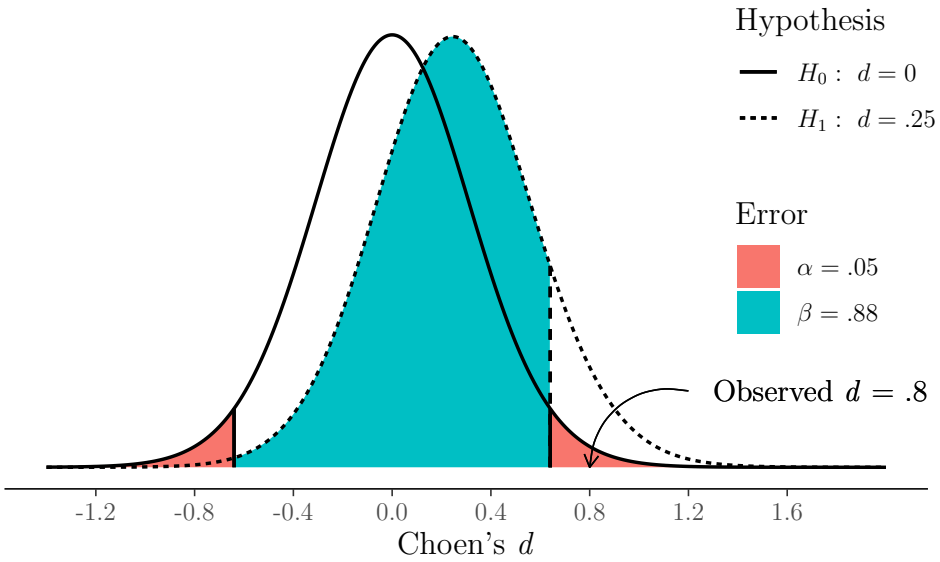

Figure 1.2: NHST comparing a treatment group and a control group ($n_{tot} = 40$), distribution under $H_0: \ d = 0$ and $H_1: \ d = .25$.

This is rather disappointing as the power level is very low. However, we may think: "*Well, we have low power, nevertheless we still have a statistically significant result. This means that our results are even more remarkable*". Unfortunately, this is a wrong deduction and it is referred to as the *"the winner's curse"* (Button et al., 2013). In the case of low power, statistically significant results are unreliable as they are almost surely an overestimation of the actual effect if an effect exists. Considering again our example, the observed effect $d = .8$ is much large than the plausible effect $H1: \ d = .25$.

Fortunately, in the last years, power analysis has become a common requirement for publication in high-quality journals. Many researchers, however, may still neglect the fundamental role of statistical power in the reliability of their results. To highlight the relation between statistical power the inferential errors related to effect size estimation, Gelman and Carlin (2014) introduced the *Design Analysis*.

1.1.3 Design Analysis

The Design Analysis enhances researchers' awareness about the consequence of conducting underpowered studies. In particular, what could pass unnoticed is that, in the case of underpowered studies, there is not only a higher probability of not rejecting the Null Hypothesis if this is false but, even more importantly, there is also a higher risk of obtaining misleading estimates in case of significant results.

While traditional power analysis has a narrow focus on statistical significance, the Design Analysis extends it to evaluate also other inferential risks related to effect size estimation (Gelman & Carlin, 2014). In particular, the Design Analysis compute:

- **Type-M error** - the predictable average overestimation of an effect that emerges as statistically significant
- **Type-S error** - the risk that a statistically significant effect is estimated in the wrong direction

Considering the previous example with a plausible effect size of $d = .25$, we would obtain a Type-M error of 3.19 (i.e., on average statistically significant results overestimate the actual effect of more than three times) and a Type-S error of 2% (i.e., two percent of statistically significant results would be in the wrong direction). At this point, we would be much more cautious in the interpretation of the results and, hopefully, we would consider the necessity to replicate the results with a much larger sample.

The Design Analysis framework will be presented in detail in the first part of the thesis. Interestingly, the Design Analysis also helps us to understand the reasons beyond the replicability crisis. In fact, psychological studies are usually characterized by heterogeneity of phenomena under investigation, noisy data, and low statistical power (Stanley et al., 2018). It is clear that, in these conditions, applying the significance filter will likely result in an overestimation of the actual effects (if the actual effects exist) that we later observed in the replicability crisis.

However, Design Analysis per se does not solve the issues related to the NHST, but it only helps to highlight its consequences. In particular, the NHST does not allow researchers to answer the question they usually are more interested in, that is evaluating the evidence in favour of their hypotheses. To overcome this limit, we need to move away from significance testing towards the evaluation of informative hypotheses using the model comparison.

1.1.4 Model Comparison

Abuse of the NHST, led to a narrow focus on testing by mechanically applying statistical techniques rather than properly evaluate the hypotheses of interest. Usually,

researchers have specific expectations or theoretical hypotheses they would like to evaluate. In this case, however, the NHST is not a suitable approach as it does not allow quantifying the evidence in favour of a hypothesis but only against it. To evaluate the evidence in favour of a hypothesis according to the data, thus, it is necessary to follow a different statistical approach, for example, the model comparison approach.

Model comparison requires first to formalize the research hypotheses into different statistical models. This is an important step as it forces researchers to define appropriate statistical models that reflect the data generative process. Moreover, it allows to specify and clarify the underlying assumptions of the hypotheses of interest.

Subsequently, it is possible to evaluate which is the most supported model among those considered according to the data. To do that different criteria are available for example the information criteria or the Bayes Factor. In particular,

- **Information Criteria** evaluate the quality of a model according to its ability to predict new data. Popular information criteria are the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

- **Bayes Factor** quantifies the relative support of the data for two competing models. In particular, within a Bayesian framework, the Bayes Factor evaluates how likely the data are under the two models a priori.

These approaches will be presented in detail in the second part of the thesis. Overall, as suggested at the beginning with the provocative sentence *"stop testing start modeling"*, we think that focusing on modeling rather than testing would enhance researchers' reasoning about statistical inference. In particular, researchers should avoid mindless testing and mechanical application of statistical techniques favouring instead the formalization of the phenomena of interest into statistical models and the evaluation of informative hypotheses. Using Luce (1988) words "[doing good research is] measuring effects, constructing substantive theories of some depth, and developing probability models and statistical procedures suited to these theories" (p.582).

1.2 Thesis Outline

1.2.1 Part I: Design Analysis

In the first part of the thesis, we offer an introduction to the Design Analysis framework considering common effect size measures and examples in psychology. In particular,

- **Chapter 2** - We introduce the elements of the Design Analysis illustrating its advantages over traditional power analysis. Differences between two independent groups are considered using Cohen's d as a measure of effect size.
- **Chapter 3** - Design Analysis is extended to the case of Pearson's correlation coefficients.
- **Chapter 4** - We present the `PRDA` R-package. This package allows performing design analysis in the case of Pearson's correlation between two variables or mean comparisons.

1.2.2 Part II: Model Comparison

In the second part of the thesis, we present the model comparison approach using the information criteria and the Bayes Factor considering two different real case applications.

- **Chapter 5** - We introduce the model comparison approach using the information criteria to evaluate the stereotype threat effects on Italian girls' mathematics performance.
- **Chapter 6** - We introduce the model comparison approach using the Bayes Factor to evaluate the relative evidence of different theoretical perspectives regarding the role of mother and father attachment.

1.2.3 Appendix and Supplemental Materials

Considering the important role of collaboration in modern research, we present in Appendix A `trackdown`, an R package offering a simple solution for collaborative writing and editing of reproducible documents (i.e. R-Markdown documents).

Moreover, in line with modern open science practices, the Supplemental Materials of each study are available online. Links are provided at the beginning of each Chapter.

1.2.4 Info Boxes

Along the thesis you will find two different boxes with special information.

Road Map

At the beginning of each chapter, we provide a brief summary. This is intended to facilitate the reading of the thesis and to help follow the logical connections between the chapters.

Round Table

Given the great value of the reviewers' comments, we decided to include a dedicated section to expand on all the relevant issues at the end of Chapter 2, 5, and 6. This is intended to create an open, constructive discussion that reminds us of the importance of analyzing everything with a critical eye. Science should always be open to constructive debating as this allows to consider new valuable perspectives opening the way for improvement.

Part I

Design Analysis

2

Enhancing Statistical Inference via Design Analysis[1]

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

> ***A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest. Remember "ATOM"*
>
> — **Wasserstein et al. (2019, p. 2)**

Road Map

In this chapter, we introduce the Design Analysis that allows researchers to evaluate the inferential risks related to effect size estimation in underpowered studies. In particular,

[1]This chapter is adapted from Altoè et al. (2020), in which I contributed to the development of the original idea, writing of the manuscript, statistical analysis and the graphical representations. Supplemental Materials available at https://osf.io/j8gsf/. Full reference:
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. https://doi.org/10.3389/fpsyg.2019.02893

in case of underpowered studies, there is not only a higher probability of not rejecting the Null Hypothesis if this is false, but, even more importantly, there is also a higher risk of obtaining misleading estimates in case of significant results. Examples considering the Cohen's d as effect size are presented.

2.1 Introduction

In the past two decades, psychological science has experienced an unprecedented replicability crisis (Ioannidis, 2005; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012) which uncovered a number of problematic issues, including the adoption of Questionable Research Practices (John et al., 2012) and Questionable Measurement Practices (Flake & Fried, 2020), the reliance on excessively small samples (Button et al., 2013), the misuse of statistical techniques (Pastore et al., 2019), and the consequent misleading interpretation and communication of research findings (Wasserstein et al., 2019).

Whereas some important reasons for the crisis are intrinsically related to psychology as a science (Chambers, 2019), which lead to a renewed recommendation to rely on strong and well-formalized theories when planning a study, the use of statistical inference undoubtedly plays a key role. Specifically, the inferential approach most widely used in psychological research, namely Null Hypothesis Significance Testing (NHST), has been strongly criticized (Gelman, 2018; Gigerenzer et al., 2004; McShane et al., 2019). As a consequence, several alternative approaches have received increasing attention, such as the use of Bayes Factors for hypothesis testing, and the use of both Frequentist and Bayesian methods to estimate the magnitude of the effect of interest with uncertainty (see J. K. Kruschke and Liddell, 2018a, for a comprehensive historical review).

In the current paper we focus on an upstream, but still neglected issue that is unrelated to the approach chosen by the researcher, namely the need for statistical reasoning, i.e., "to reason about data, variation and chance" (Moore, 1998, p. 1253), during all phases of an empirical study. Our work was inspired by the famous statistician Ronald Fisher (1890 -1962), who stated that "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of" (Fisher, 1938, p. 17). Indeed, we argue that too often statistical inference is seen as an isolated procedure, which is limited to the analysis of data that have already been collected. In particular, we emphasize the non-trivial importance of making statistical considerations at the onset of a research project. Furthermore, we stress that although Fisher has ironically defined them a "post mortem examination", appropriate evaluations of published results can provide a relevant contribution to the progress of (psychological) science. The ultimate goal of this paper is to in-

crease researchers' awareness by promoting active engagement when designing their research.

To achieve this goal, we build on and further develop an idea proposed by Gelman and Carlin (2014) called "prospective and retrospective design analysis", which is virtually absent in current research practice. Specifically, to illustrate the benefits of design analysis to the widest possible audience, we use a familiar example in psychology where the researcher is interested in analyzing the differences between two independent groups considering Cohen's d (Cohen, 1988) as an effect size measure.

In brief, the term *design analysis* has been proposed by Gelman and Carlin (2014) as a broader definition of power analysis, a concept that in the statistical literature traditionally indicates the determination of an appropriate sample size given prespecified levels of Type I and Type II errors and a "plausible effects size" (Gigerenzer et al., 2004). Indeed, a comprehensive design analysis should also explicitly consider other inferential risks, including the exaggeration ratio (Type M error, i.e., the predictable average overestimation of an effect that emerges as statistically significant) and the sign error (Type S error, i.e., the risk that a statistically significant effect is estimated in the wrong direction). Notably, the estimation of these errors will require an effort from psychologists to introduce their expert knowledge and hypothesize what could be considered a "plausible effect size". As we will see later, a key aspect of design analysis is that it can be usefully carried out both in the planning phase of a study (i.e., prospective design analysis) and for the evaluation of studies that have already been conducted (i.e., retrospective design analysis).

Although the idea of design analysis could be developed within different inferential statistical approaches (e.g., Frequentist and Bayesian), in this paper we will rely on the Neyman-Pearson (N-P) approach (Pearson & Neyman, 1928) as opposed to the widely used NHST. The rationale for this choice is that, in addition to other strengths, the N-P approach includes formalization of the *Null Hypothesis* (i.e., the absence of an effect) like NHST, but also an explicit formalization of the *Alternative Hypothesis* (i.e., the magnitude of the expected effect). For a more comprehensive description of the difference between N-P and NHST approaches, we refer the reader to Gigerenzer and colleagues (2004).

The remainder of this paper is structured as follows. In the next paragraphs, we will briefly review the main consequences of underpowered studies, discuss two relevant misconceptions concerning the interpretation of statistically significant results, and present a theoretical framework for design analysis including some clarifications regarding the concept of "plausible effect size". In Section 2.2, through familiar examples within psychological research, the benefits of prospective and retrospective design analysis will be highlighted. Subsequently, in Section 2.3, a real case study

will be presented and analyzed. Finally, in Section 2.4, we will summarize potentials, further developments, and limitations of our proposal.

2.1.1 The Consequences of Underpowered Studies in Psychology

In 1962, Cohen raised attention towards a problem affecting psychological research that is still very much alive today (Cohen, 1962). Researchers seemed to ignore the statistical power of their studies - which is not considered in NHST (Gigerenzer et al., 2004) - with severe consequences for the robustness of their research findings. In the N-P 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. One of the problems with underpowered studies is that the probability of finding an effect, if it actually exists, is low. More importantly, if a statistically significant result (i.e., "in general", when the observed $p - value$ is less than .05, and consequently H_0 is rejected; see, Wasserstein et al., 2019) is obtained in an underpowered study, the effect size associated with the observed $p - value$ might be "too big to be true" (Button et al., 2013; Gelman & Carlin, 2014).

This inflation of effect sizes can be seen when examining results of replication projects, which are usually planned to have higher power than the original studies. For example, the Open Science Collaboration (2015, pp. 4-5) reported that "Overall, original study effect sizes ($M = 0.403$, $SD = 0.188$) were reliably larger than replication effect sizes ($M = 0.197$, $SD = 0.257$)", and in the Social Science Replication Project (Camerer et al., 2018, p. 637), "the effect size of the replication was on average about 50% of the original effect size". These considerations contributed to the introduction in the literature of the term "decline effect", defined as "the notion that science routinely observes effect sizes decrease over repeated replications for reasons that are still not well understood" (Schooler, 2014, p. 579).

Given that underpowered studies are widespread in psychology (Cohen, 1962; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989), the shrinkage of effect-sizes in replications could be partially explained by the fallacy of "what does not kill statistical significance makes it stronger" (Loken & Gelman, 2017) and by the trap of the "winner's curse" (Button et al., 2013).

2.1.2 The "Winner's Curse" Trap

When a statistically significant result is obtained in an underpowered study (e.g., power = 40%), in spite of the low probability of this event to happen, the result might be seen as even more remarkable. In fact, the researcher might think: "If obtaining a statistically significant result is such a rare event, and in my experiment

I obtained a statistically significant result, it must be a strong one". This is called the "what does not kill statistical significance makes it stronger" fallacy (Loken & Gelman, 2017). The reason why this is a fallacy lies in the fact that it is possible to obtain statistical significance because of many other factors different from the presence of a real effect. The researcher degrees of freedom, large measurement errors, and small sample sizes all contribute to create noise in the data, therefore inflating the perhaps true, but small underlying effect. Then, if the procedure used to analyze those data is only focused on a threshold (like in NHST, with the conventional significance level of .05), the noise in the data allows to pass this threshold.

In these situations, the apparent win in terms of obtaining a statistically significant result is actually a loss, in that the "the lucky" scientist who makes a discovery is cursed by finding an inflated estimate of that effect (Button et al., 2013). This is called the "Winner's curse", and Figure 2.1 shows an example. In this hypothetical situation, the researcher is interested in studying an effect that can plausibly be of small dimensions, e.g. Cohen's d of .20. If s/he decides to compare two groups on the outcome variable of interest, using 33 participants per group (and performing a two-tailed test), s/he will never be able to simultaneously reject H_0 and find an effect close to what it is plausible in that research field (i.e., .20). In fact, in this underpowered study (i.e., based on a d of .20, the actual power is only 12%) all the effects falling in the "rejection regions" are higher than .49 or smaller than -.49, and .20 falls in the region where the decision rules state that you cannot reject H_0 under the NHST approach, nor can you accept H_0 under the N-P approach.

2.1.3 Beyond Power: The Design Analysis

As we saw in the previous example, relying solely on the statistical significance of a result can lead to completely misleading conclusions. Indeed, researchers should take into account other relevant information, such as the hypothesized "plausible effect size" and the consequent power of the study. Furthermore, to assist researchers with evaluating the results of a study in a more comprehensive way, Gelman and Carlin (2014) suggested to consider other two relevant types of errors in addition to the traditional Type I and Type II errors, namely Type M and Type S error (see also, Gelman & Tuerlinckx, 2000; Lu et al., 2019). Specifically, Type M [magnitude] error or *exageration ratio* can be viewed as the the expected average overestimation of an effect that emerges as statistically significant, whereas Type S [sign] error can be viewed as the probability of obtaining a statistically significant result in the opposite direction with respect to the sign of the hypothesized plausible effect size.

Based on this consideration, Gelman and Carlin (2014) proposed the term *"de-*

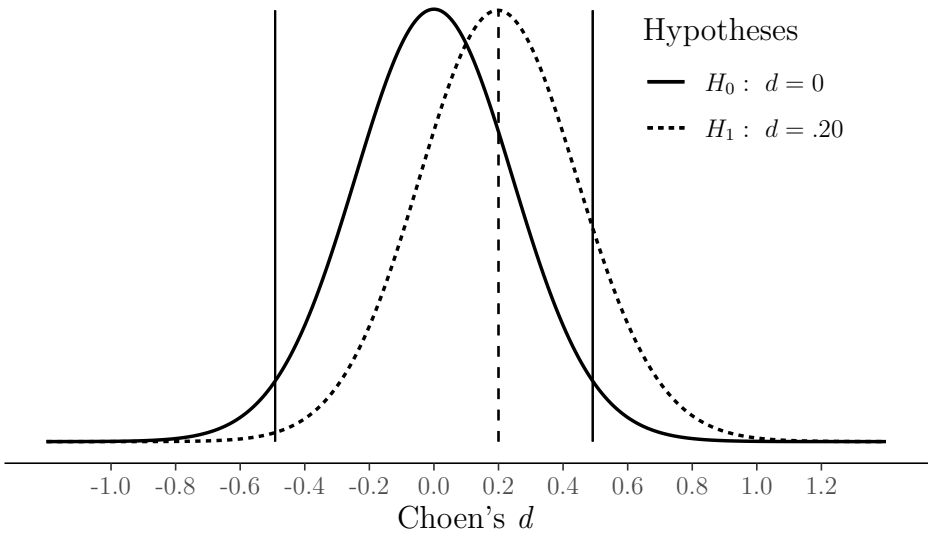

Figure 2.1: The Winner's Curse. Hypothetical study where the plausible true effect size is small (Cohen's $d = .20$) and a two tailed independent samples t-test is performed with 33 people per group. In order to reject H_0, the researcher has to overestimate the underlying true effect which is indicated by the dashed vertical line. Note: the rejection regions of H_0, given a significance level of .05, lie outside the vertical black lines

sign analysis" to broadly identify the analysis of studies' properties, such as their statistical power, Type M and Type S error. Moreover, as we shall see in the next paragraph, in design analysis particular emphasis is given on the elicitation and formalization of what can be considered a *plausible effect size* (see also paragraph 2.1.4) for the study of interest. In this regard, it is important to make a clarification. Although Gelman and Carlin (2014) developed design analysis relying on an unstandardized effect size measure (i.e., the difference between two means), in this paper we have adapted their method to deal with Cohen's d, a standardized measure of effect size that is more commonly used in psychology.

Given these premises, the steps to perform design analysis using Cohen's d as a measure of effect size can be summarized in three steps:

1. A plausible effect size for the study of interest needs to be identified. Rather than focusing on data at hand or on noisy estimates of a single pilot study, the formalization of a plausible effect size should be based on an extensive theoretical literature review and/or on meta-analyses. Moreover, specific tools (see for example O'Hagan, 2019; Zandonella Callegher et al., 2019; Zondervan-Zwijnenburg et al., 2017) that allow to incorporate expert knowledge can also

be considered to increase the validity of the plausible effect size elicitation process.[2]

2. Based on the experimental design of the study of interest (in our case, a comparison between two independent groups), a large number of simulations (i.e., 100,000) will be performed according to the identified plausible effect size. This procedure serves to provide information about what to expect if the experiment was replicated an infinite number of times assuming the pre-identified plausible effect as true.

3. Given a fixed level of Type I error (e.g., .05), power, type M and type S error will be calculated. Specifically, power will be estimated as the ratio between the number of obtained significant results and the number of replicates (i.e., the higher the power, the higher the probability to detect the plausible effect). Type M error will be estimated as the ratio between the mean of the absolute values of the statistically significant replicated effect sizes and the plausible effect size. In this case, larger values indicate an expected large overestimation of the plausible effect size. Type S error will be the ratio between the number of significant results with opposite sign with regard to the plausible effect size and the the total number of significant results. Put in other terms, type S error estimates the probability of obtaining a significant result in the wrong direction.

Although the procedure may seem complex to implement, at the link https://osf.io/wqd7b/ we made available some easy-to-use R functions that allow to perform different types of design analysis for less experienced users. The same functions will also be used in the examples and application presented in this paper.

To get a first idea of the benefits of design analysis, let us re-analyze the hypothetical study presented in Figure 2.1. Specifically, given a plausible effect size equal to $d = .20$ and a sample size of 33 participants per group, a design analysis will highlight the following information: power = 12%, Type M error = 3.10, and Type S error = 2%. Despite the low power, which shows that the study has only a 12% probability to detect the plausible effect size, type M error explicitly indicates that the expected overestimate of a result that will emerge as statistically significant is around 3 times the plausible effect. Furthermore, given a Type S error of 2%, there is also a non negligible probability of obtaining a significant result in the wrong direction. Overall, the results of design analysis clearly tell the researcher

[2]To obtain a more comprehensive picture of the inferential risks associated with their study, we suggest researchers to inspect different scenarios according to different plausible effect sizes and thus to perform more than one design analysis (see for example our application to a real case study in Section 2.3).

that the study of interest could provide very poor support to both the existence and non-existence of a plausible effect size.

Another advantage of design analysis, which will be better explored in the following sections, is that it can be effectively used in the planning phase of a study, i.e., *prospective design analysis*, as well as in the evaluation of already obtained study results, i.e., *retrospective design analysis*. For example, in prospective design analysis, considerations concerning power, Type M, and Type S error could assist researchers to decide the appropriate sample size for detecting the effect of interest (if it actually exists). In a retrospective design analysis, power, Type M and Type S error (always calculated using the theoretically plausible effect size) can be used to obtain information about the extent to which the results of the study could be exaggerated and/or in the wrong direction. Most importantly, we believe that, engaging in a retrospective design analysis helps researchers to recognize the role of uncertainty and to make more reasonable statistical claims, especially in those cases at risk of falling in the aforementioned "Winner's Curse" trap.

In conclusion, it is important to note that whatever the type of design analysis chosen (prospective or retrospective), the relationships between power, type M error, and type S error are the same. For illustrative purposes, these relationships are graphically displayed as a function of sample size in Figure 2.2. A medium-to-small effect of $d = .35$ (i.e., a reasonable average effect size for a psychological study in the absence of other relevant information, see also Section 2.3) was considered as a plausible effect size, and Type I error was set at .05.

As expected, power increases as sample size increases. Moreover, type M and type S error decrease as the size of the sample increases, with the latter showing a much steeper decrease.

From an applied perspective, issues with type M and S errors emerge with underpowered studies which are very common in psychological research. Indeed, as can be seen in Figure 2.2, for a power of 40% (obtained with 50 participants per group), type M error reaches the worrisome value of 1.55; for a power around 10% (i.e., with 10 participants per group), even type S error becomes relevant (around 5%).

2.1.4 What Does "Plausible Effect Size" Mean?

> *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)**

The main and most difficult point rests on deciding what could be considered a "plausible effect size". Although this might seem complex, usually studies are not

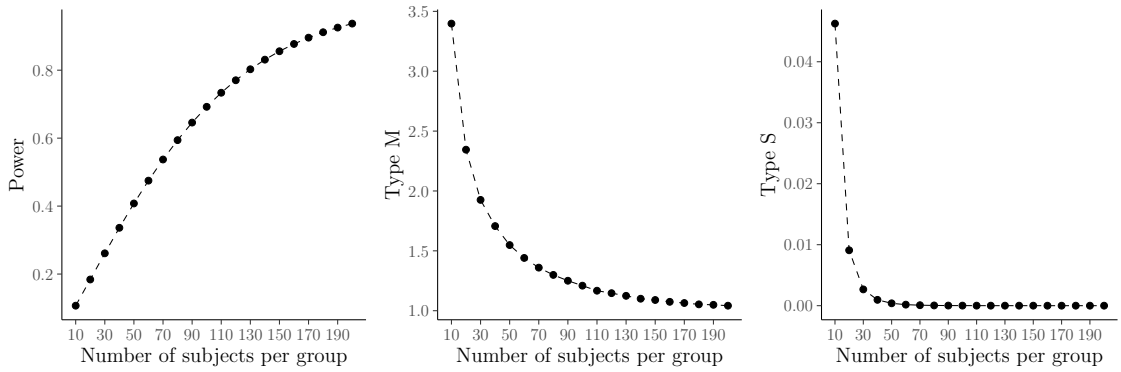

Figure 2.2: Relationship between sample size and Power, Type M and Type S for a Cohen's d of .35 in an independent samples t-test. Type I error is set at .05.

developed in a void. Hypotheses are derived from theories that, if appropriately formalized in statistical terms, will increase the validity of the inferential process. Furthermore, researchers are commonly interested in knowing the size and direction of effects; as shown above, this corresponds to control for Type M [magnitude] error and type S [sign] error.

From an epistemological perspective, J. Kruschke (2013) suggests an interesting distinction between *strong theories* and *weak theories*. Strong theories are those that try to make precise predictions and could be, in principle, more easily disconfirmed. For example, a strong theory could hypothesize a medium-sized positive correlation between two variables. In contrast, weak theories make broader predictions, such as the hypothesis that two variables are correlated without specifying the strength and direction of the correlation (Dienes, 2008). The former type allows many more research findings to disconfirm the hypothesis, whereas the latter type allows only the result of no correlation to disconfirm it. Specifically, following Karl Popper (1902–1994), it could be argued that theories explaining virtually everything and being hard to disconfirm risk to be out of the realm of science[3]. Thus, scientific theories should provide at least a hint on the effect that is expected to be observed.

A challenging point is to establish the dimension of this effect. It might seem paradoxical that the researcher has to provide an estimate of the effect size before running the experiment, given that s/he will conduct the study exactly with the aim of finding what that estimate is. However, strong theories should allow to make such predictions, and the way in which science accumulates should provide increasing precision to these predictions.

In practice, it might be undesirable to simply take the estimate found in a pilot study or from a single previous study published in the literature as a "plausible

[3]For a recent discussion about theory crisis in psychology see Eronen and Bringmann (2021)

effect size". In fact, the plausible effect size refers to what could be approximately the true value of the parameter in the population, whereas the results of pilots or single studies (especially if underpowered) are noisy estimates of that parameter.

In line with Gelman and Carlin (2014), we suggest to use information outside the data at hand, such as literature reviews and/or meta-analyses taking into account issues concerning publication bias (Borenstein et al., 2009). Moreover, as stated in the previous paragraph, promising procedures to elicit and formalize expert knowledge should also be considered. It is important to note that, whatever the procedures, all assumptions that will lead to the identification of a plausible effect size must be communicated in a transparent manner, thus increasing the information provided by a study and ensuring more reasonable statistical claims related to the obtained results, whether they are significant or not.

As we have seen, the identification of a plausible effect size (or a series of plausible effect sizes to explore different scenarios) requires a big effort from the researcher. Indeed, we believe that this kind of reasoning can make a substantial contribution to the planning of robust and replicable studies, as well as to the efficient evaluation of obtained research findings.

To conclude, we leave the reader with the following question: "All other conditions being equal, if you had to evaluate two studies of the same phenomenon, the first based on a formalization of the expected plausible effect sizes of interest that is as accurate as possible, and the second one in which the size of the effects of interest was not taken into account, the findings of which study would you believe the most?" (van de Schoot, 2019).

2.2 Prospective and Retrospective Design Analysis

To highlight the benefits of design analysis, and to familiarize with the concepts of Type M and Type S errors, we will start from a simple example that is well known in psychological research, i.e., the comparison between the means of two independent groups[4].

In particular, the goal of our hypothetical case study is to evaluate the differences between two treatments that aim to improve a cognitive ability called Y. Both treatments have the same cost, but the first is innovative, whereas the second is traditional. To this end, the researchers recruit a sample of participants who are homogeneous with respect to prespecified relevant study variables (i.e., age, IQ ...). Next, they randomly assign each participant to one of the two conditions (i.e.,

[4]We remind the reader that source code of ad-hoc R functions used in the paper is available at the link https://osf.io/wqd7b/

innovative vs traditional treatment). After the treatment phase is completed, the means of the two groups are compared.

2.2.1 Prospective Design Analysis

Before collecting data, the researchers decide to plan the appropriate sample size to test their hypotheses, namely that there is a difference between the means of $G1$ (the group to which the innovative treatment is administered) and $G2$ (the group to which the traditional treatment is administered) *vs* there is no difference.

After an extensive literature review concerning studies theoretically comparable to their own, the researchers decide that a first reasonable effect size for the difference between the innovative and the traditional treatment could be considered equal to a Cohen's d of .30. Due to the possible presence of publication bias (Borenstein et al., 2009), which could lead to an overestimation of the effects of published studies, the researchers decide to be more conservative about the estimate of their plausible effect size. Thus, they decide to consider a Cohen's d of .25. Eventually, all researchers agree that a Cohen's d of .25 could also represent a clinically relevant effect in order to support the greater efficacy of the innovative treatment.

Based on the above considerations, the researchers start to plan the sample size for their study. First, they fix the Type I error at .05 and - based on commonly accepted suggestions from the psychological literature - fix the power at .80. Furthermore, to explicitly evaluate the inferential risks connected to their choices they calculate the associated Type M and Type S errors.

Using our `R` function `design_analysis`, they obtain the following results:

```
design_analysis(d=.25, power=.80)

##     d power   n typeS typeM
##  0.25  0.80 252  0.00  1.13
```

Based on the results, to achieve a power of .80, a sample size of 252 for each group is needed (i.e., total sample size = 504). With this sample size, the risk of obtaining a statistically significant result in the wrong direction (Type S error) is practically 0 and the expected exaggeration ratio (Type M error) is 1.13. In other words, the expected overestimation related to effects that will emerge as statistically significant will be around 13% of the hypothesized plausible effect size.

Although satisfied in terms of expected type S and type M risks, the researchers are concerned about the economic feasibility of recruiting such a "large" number of subjects. After a long discussion, they decide to explore which inferential risks would result for a lower level of power, namely 60%[5].

[5]Specifically, we agree with Gelman (2019a) that an 80% level of power should not be used as an

Using the function `design_analysis`

```
design_analysis(d=.25, power=.60)

##     d power   n typeS typeM
##  0.25  0.60 158  0.00  1.30
```

they discover that: 1) the overall required sample size is considerably smaller (from 504 to $316 = 158 \times 2$), thus increasing economic feasibility of the study; 2) the Type S error remains negligible (0%) ; 3) the exaggeration ratio considerably increases (from 1.13 to 1.30); thus, an effect that will emerge as statistically significant will be on average 130% of the hypothesized plausible effect size.

The researchers now need to make a decision. Even though, from a merely statistical point of view, the optimal choice would be to consider a power of 80%, other relevant aspects must be evaluated, such as the possibility to obtain additional funding, the practical implications of an expected overestimation of the plausible effect size, and the phase of the study (i.e., preliminary/exploratory, intermediate or final/confirmatory).

Whatever the decision, the researchers have to be aware of the inferential risks related to their choice. Moreover, when presenting the results, they have to be transparent and clear in communicating such risks, thus highlighting the uncertainty associated with their conclusions.

2.2.2 Retrospective Design Analysis

To illustrate the usefulness of retrospective design analysis, we refer to the example presented in the previous paragraph. However, we introduce three new scenarios which can be considered as representative of what commonly occurs during the research process:

- **Scenario 1 (S1): Evaluating sample size based on a single published study.**[6]
  Imagine that the researchers decide to plan their sample size based on a single published study in the phase of formalizing a plausible effect size, either because the published study presents relevant similarities with their own study, or because there are no other published studies available.

automatic routine, and that requirements of 80% power could encourage researchers to exaggerate their effect sizes when planning sample size.

[6]Even though in this paper we strongly recommend not to plan sample size based on a single study, we propose this example to further emphasize the inferential risks associated with the information provided by a single underpowered study.

Question: What type of inferential risks can be associated with this decision?
Issues: Using a single study as a reference point without considering other sources (e.g., theoretical framework, experts' opinion, or a meta-analysis), especially when the study has a low sample size and/or the effect of interest is small, can lead to use an excessively optimistic estimate of the effect size to plan an appropriate sample size (Gelman & Carlin, 2014).

- **Scenario 2 (S2): Difficulty in recruiting the planned number of research participants.**
  Imagine that, due to unforeseen difficulties (e.g., insufficient funding), the researchers are not able to recruit the pre-planned number of participants as defined based on prospective design analysis.
  Question: How to evaluate the inferential risks associated with the new reduced sample size? How to communicate the obtained results?
  Issues: Researchers are often tempted to evaluate the results of their study based on the observed effect size. This procedure, known as "post-hoc power analysis", has been strongly criticized and many statistical papers explicitly advise against its use (see for example, Gelman, 2019a; S. Goodman & Berlin, 1994). Indeed, to evaluate the information provided by the obtained results, researchers should use the a priori plausible effect size, i.e., the one formalized before collecting their data.

- **Scenario 3 (S3): No prospective design analysis because the number of participants is constrained.**
  Imagine the number of participants involved in the study have specific characteristics which make it impossible to yield a large sample size, or that the type of treatment is particularly expensive and therefore it cannot be tested on a large sample. In this case, the only possibility is to recruit the largest possible number of participants.
  Question: What level of scientific quality can be provided by the results?
  Issues: Although study's results can provide a useful contribution to the field, there are several associated inferential risks that the researchers need to communicate in a transparent and constructive way.

As we will see below, retrospective design analysis can be a useful tool to deal with the questions and the issues raised across all three scenarios.

For the sake of simplicity and without loss of generalizability, suppose that in each of the three scenarios the researchers obtained the same results (see Table 2.1).

At a first glance, the results indicate a statistically significant difference in favor of the innovative treatment (see Table 2.1), with a large effect size (i.e., $d = .90$). However, the 95% confidence interval for Cohen's d is extremely wide, suggesting

Table 2.1: Comparison of the cognitive skill Y between the two groups.

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

that both medium-small (i.e., $d = .38$) and very large (i.e., $d = 1.43$) effects are consistent with the observed data.

A closer look indicates that the estimated effect size seems too large when compared with the initial guess of the researchers (i.e. $d = .25$). Furthermore, an estimated d of .90 seems, in general, implausibly large for a difference between two cognitive treatments. The latter interpretation seems to be also supported by the fact that the hypothesized plausible effect size is not even included in the estimated confidence interval. Overall, in order to prevent the aforementioned "Winner's Curse" and "What Does Not Kill Statistical Significance Makes It Stronger" heuristics, results must be evaluated and eventually communicated with caution and skepticism.

To obtain a clearer picture of the inferential risks associated with the observed results, we can perform retrospective design analysis using $d = .25$ as plausible effect size and 31 participants per group as sample size:

```
design_analysis(n=31, d=.25)

##     d  n power typeS typeM
##  0.25 31  0.16  0.01  2.60
```

As can be seen, the power is markedly low (i.e, only 16%) and the Type M error even suggests an expected overestimation around two and a half times the plausible effect size. Lastly, the Type S error, although small, indicates a 1% risk of obtaining a significant result in the wrong direction (i.e., the traditional treatment is better than the innovative treatment). Let's see how this information could be helpful to deal with the three presented scenarios.

In S1, the researchers took a single noisy estimate as the plausible effect size from a study that found a "big" effect size (e.g., 0.90). The retrospective design analysis shows what happens if the plausible effect size is, in reality, much smaller (i.e., 0.25). Specifically, given the low power and the high level of Type M error, researchers should definitely abandon the idea of planning their sample size based on a single published study. Furthermore, issues regarding the presence of Questionable Research Practices (Arrison, 2014; John et al., 2012) and Questionable Measurement Practices (Flake & Fried, 2020) in the considered published study

must be at least explored. From an applied perspective, researchers should continue with a more comprehensive literature review and/or consider the opportunity to use an effect size elicitation procedure based on experts' knowledge (O'Hagan, 2019; Zondervan-Zwijnenburg et al., 2017).

In S2, to check the robustness of their results, researchers might initially be tempted to conduct a power analysis based on their observed effect size ($d = .90$). Acting in this way, they would obtain a completely misleading post-hoc power of 94%. In contrast, the results of retrospective design analysis based on the a-priori plausible effect size ($d = .25$) highlight the high level of inferential risks related to the observed results. From an applicative perspective, researchers should be very skeptical about their observed results. A first option could be to replicate the study on an independent sample, perhaps asking for help from other colleagues in the field. In this case, the effort to recruit a larger sample could be well-justified based on the retrospective design analysis.

In S3, given the low power and the high level of Type M error, results should be presented as merely descriptive by clearly explaining the uncertainty that characterizes them. Researchers should first reflect on the possibility of introducing improvements to the study protocol (i.e., improving the reliability of the study variables). As a last option, if improvements are not considered feasible, the researchers might consider not continuing their study.

Despite its advantages, we need to emphasize that design analysis should not be used as an *automatic problem solver machine*: "Let's pull out an effect size ... let me see the correct sample size for my experiment". In other words, to obtain reliable scientific conclusions there is no "free lunch". Rather, psychologists and statisticians have to work together, case by case, to obtain a reasonable effect size formalization and to evaluate the associated inferential risks. Furthermore, researchers are encouraged to explore different scenarios via sensitivity analysis (see, Section 2.3) to better justify and optimize their choices.

2.3 An Illustrative Application to a Case Study

To illustrate how design analysis could enhance inference in psychological research, we will consider a real case study. Specifically, we will focus on Study 2 of the published paper "A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action" (Kay et al., 2014).

The paper presents 5 studies arising from findings showing that human beings have a natural tendency to perceive structure in the surrounding world. Various social psychology theories propose plausible explanations which share a similar assumption that had never been tested before, that is, perceiving a structured world

could increase people's willingness to make efforts and sacrifices towards their own goals. In Study 2, the authors decided to test this hypothesis by randomly assigning participants to two different conditions differing in the type of text they had to read. In the "random" condition, the text conveyed the idea that natural phenomena are unpredictable and random, whereas in the "structure" condition the phenomena were described as predictable and systematic. The outcome measure was the willingness to work towards a goal that each participant chose as their "most important". The expected result was that participants in the "structure" condition would report a higher score in the measure of goal directed behavior than those in the "random" condition.

Prospective design analysis

As we saw in the previous paragraphs, before collecting data is fundamental to plan an appropriate sample size via prospective design analysis. In this case, given the relative novelty of Study 2, is hard to identify a single plausible value for the size of the effect of interest. Rather, it seems more reasonable to explore different scenarios according to different plausible effect sizes and power levels. We will start with a minimum d of .20, so that the study is planned to detect at least a "small" effect size. If the final results do not reach statistical significance, the researchers could conclude that it is unlikely that the true effect is equal to or greater than .20, and eventually decide whether it is worth replicating the study, perhaps by modifying their protocol.

As the most plausible effect size, we will consider $d = .35$, which could be considered – at least in our opinion - a typical average level to test a hypothesis in psychological research in the absence of informative external sources (see for example the results reported in Open Science Collaboration, 2015)[7]. As an extrema ratio, we will include also a d of .5, which in the words of Jacob Cohen can be referred to as "differences that are large enough to be visible to the naked eye" (see Cohen, 1988, p.26), and that, given the experiment under investigation, could be viewed as an extremely optimistic guess. Finally, to take issues concerning the feasibility of the study into account, we will also consider two levels of power, namely 80% and 60%.

[7]In the Open Science Collaboration (2015), the authors conducted replications of 100 experimental and correlational studies published in three psychology journals using high powered designs and original materials when possible. They found an average effect size of $r = .197$, (i.e. $d = .41$). Given the heterogeneity of the 100 studies, we propose to use a more conservative value as being representative of a typical average effect in psychology. Overall, it should be noted that all the pre-specified values of d, albeit plausible, are not based on a thorough theoretical revision and/or on the formalized knowledge of experts in the field. Indeed, an appropriate use of the latter two external sources would undoubtedly contribute to produce more reliable results, but discussion of these strategies is beyond the scope of this paper.

Overall, our "sensitivity" prospective design analysis (see Table 2.2) suggests that the sample size chosen by the authors ($n = 67$) is inadequate. Indeed, even in the least reasonable scenario ($d = .50$, power = .60), a minimum of 80 participants is required. Furthermore, is should be noted, that the associated Type M error is considerably high, i.e. 131%, signaling a high risk of overestimating the plausible effect.

Table 2.2: Sample size, Type M and Type S error by power and plausible effect size. Type I error is fixed at .05

Power	**Cohen's** d	n **(per sample)**	**Total** n	**Type** M **error**	**Type** S **error**
0.80	0.20	392	784	1.13	0.00
	0.35	130	260		
	0.50	64	128		
0.60	0.20	244	488	1.30	0.00
	0.35	80	160		
	0.50	40	80		

A good compromise could be to consider the second scenario ($d = .35$, power = .80), which requires a total sample size of 260, guaranteeing optimal control of the Type M error. After conducting the study with this sample size, a significant result would lead to accept the researcher's hypothesis, while a non-significant result would indicate that if an effect exists, it will presumably be less than .35. Whatever the result, the researchers could eventually present their findings in a transparent and informative way. In any case, the results could be used to improve scientific progress. As an example, other researchers could fruitfully use the observed results as a starting point for a replication study.

Retrospective design analysis

Let us now evaluate Study 2 from a retrospective point of view. Based on their results ($M_{\text{structure}} = 5.26$, $SD_{\text{structure}} = 0.88$, $M_{\text{random}} = 4.72$, $SD_{\text{random}} = 1.32$, $n_{\text{total}} = 67$; $t(65) = 2.00$, $p = .05$, Cohen's $d = 0.50$)[8], the authors concluded that "participants in the structure condition reported higher willingness to expend effort and make sacrifices to pursue their goal compared to participants in the random condition." Kay et al. (2014, p. 487), thus supporting their initial hypothesis.

[8]The authors reported only the total sample size ($n = 67$). Since participants were randomly assigned to each of the two experimental conditions, in the following we will assume, without loss of generalizability, that 34 participants were assigned to the "structure" condition, and 33 to the "random" condition.

To evaluate the inferential risks associated with this conclusion, we now run a sensitivity retrospective design analysis on the pre-identified plausible effect sizes (i.e., $d = .20$, $d = .35$, $d = .50$).

In line with the results emerging from prospective analysis, the retrospective design analysis indicates that the sample size used in Study 2 exposes to high inferential risks. In fact, both for a plausible effect of $d = .20$ (power = 12%, type $M = 3.09$, type $S = 0\%$) and for a plausible effect of $d = .35$ (power = 30%, type $M = 1.86$, type $S = 1\%$), the power is very low and the Type M error reaches worrying levels. For a d of .50 (chosen on the basis of plausible effects and not based on the results observed in Study 2), the Type M error is 1.39, indicating an expected overestimate of 39%. Furthermore, the power is 51%, suggesting that if we replicate the study on a new sample with the same number of participants, the probability of obtaining a significant result will be around the chance level.

In summary, our retrospective design analysis indicates that, although statistically significant the results of Study 2 are inadequate to support the authors' conclusions.

As mentioned at the beginning of this paragraph, the Study 2 of Kay et al. (2014) was selected for illustrative purposes and for a constructive perspective. For a more comprehensive picture, we invite interested readers to consult the "Many Labs 2 project" (Klein et al., 2018), which showed that with a large sample size ($n = 6506$) the original conclusion of Study 2 cannot be supported (i.e., $t(6498.63) = -0.94$, $p = .35$, $d = -0.02$, $95\%CI = [-0.07, 0.03]$), as well as the subsequent response of the original authors (Laurin et al., 2018).

2.4 Discussion and Conclusions

In psychological research, statistical inference is often viewed as an isolated procedure which limits itself to the analysis of data that have already been collected. In this paper, we argue that statistical reasoning is necessary both at the planning stage and when interpreting the results of a research project. To illustrate this concept, we built on and further developed Gelman and Carlin's (2014) idea of "prospective and retrospective design analysis".

In line with recent recommendations (Cumming, 2014), design analysis involves an in-depth reasoning on what could be considered as a plausible effect size within the study of interest. Specifically, rather than focusing on a single pilot or published study, we underlined the importance of using information outside the data at hand, such as extensive literature reviews and meta-analytic studies taking issues related to publication bias into account. Furthermore, we introduced the potentials of elicitation of expert knowledge procedures (see for example O'Hagan, 2019; Zondervan-Zwijnenburg et al., 2017). Even though these procedures are still underutilized

in psychology, they could provide a relevant contribution to the formalization of research hypotheses.

Moving beyond the simplistic and often misleading distinction between significant and non significant results, design analysis allows researchers to quantify, consider, and explicitly communicate two relevant risks associated with their inference, namely exaggeration ratio (Type M error) and sign error (Type S error). As illustrated in the paper, the evaluation of these risks is particularly relevant in studies which investigate small effect sizes in the presence of high levels of intra- and inter-individual variability, with a limited sample size – a situation that is quite common in psychological research.

Another important aspect of design analysis is that it can be usefully carried out both in the planning phase of a study (i.e., prospective design analysis) and to evaluate studies that have already been conducted (i.e., retrospective design analysis), reminding researchers that the process of statistical inference should start before data collection and does not end when the results are obtained. In addition, design analysis contributes to have a more comprehensive and informative picture of the research findings through the exploration of different scenarios according to different plausible formalizations of the effect of interests.

To familiarize the reader with the concept of design analysis, we included several examples and an application to a real case study. Moreover, to allow researchers to use all the illustrated methods with their own data, we also provided two easy-to-use `R` functions which are available at the Open Science Framework (OSF) at the link https://osf.io/wqd7b/.

For the sake of simplicity, in this paper we limited our consideration to Cohen's d as an effect size measure within a Frequentist approach. However, the concept of design analysis could be extended to more complex cases and to other statistical approaches. For example, our `R` functions could be directly adapted to other effect size measures, such as Hedges' g, Odds Ratio, η^2 and R^2.

Also, it is important to note that our considerations regarding design analysis could be fruitfully extended to the increasingly used Bayesian methods. Indeed, our proposed method to formalize uncertainty via probability distributions finds its natural extension in the concept of Bayesian prior. Specifically, design analysis could be useful to evaluate the properties and highlight the inferential risks (such as type M and type S errors) associated with the use of Bayes Factors and parameter estimation with credible Bayesian intervals.

In sum, even though a design analysis requires big effort, we believe that it has the potential to contribute to planning more robust studies and promoting better interpretation of research findings. More generally, design analysis and its associated way of reasoning helps researchers to keep in mind the inspiring quote presented at the beginning of this paper regarding the use of statistical inference:

"Remember ATOM".

Round Table

1. The concept of "*plausible effect size*" is sometimes used interchangeably to the concept of "*population effect size*". In particular, the use of NHST in underpowered studies (plus some form of publication bias that prevents null results to be available to the scientific community) leads to overestimating effect sizes not just compared to what could be considered plausible, but compared to what is true in the population of interest (which is a worse problem). Then of course since the population effect size is unknown, the plausible effect size can be used as a proxy in power and design analysis, but this should be perhaps discussed more clearly.

   Answer: We agree with the reviewer that using the terms "*plausible effect size*" interchangeably to the concept of "*population effect size*", could be confusing. As suggested by the reviewer, the plausible effect size is used as a proxy for the population effect size in power and design analysis. However, since "*the population effect sizes is unknown*", we will not know whether this is a good approximation or not. As the reviewer underlined, this could be problematic since researchers could easily overestimate the population effect sizes due to several issues in the literature (e.g., underpowered studies and publication bias). For these reasons, we underlined the importance of critically evaluating the literature taking into account all available information. As an alternative, researchers could define the plausible effect size according to other requirements, for example considering the minimum effect size of interest. Design analysis can be straightforwardly extended to these cases as well. In the framework in which we introduced the design analysis, however, plausible effect size refers to what could be approximately the true value of the parameter in the population.

2. In Section 2.2.2, three specific scenarios are presented to exemplify the retrospective design analysis, giving practical suggestions in each scenario. However, a general description of the retrospective design analysis and its utility, before presenting the scenarios, could be useful in my opinion. In general, whereas I see the value of the prospective design analysis, the dissertation could do more to convince the reader about the utility of the retrospective design analysis. Scenario S1, in particular, puzzles me a lot. In this scenario, a research team plans a sample size based on a single published study (e.g., because it is the only study available). Suppose that this effect size is d = .72, since they collect N = 62 participants for a two-tailed t-test (as in Table 2.1) and I assume that they used alpha = .5 and beta = .20. After the study, they obtain as a result an effect d = .90. According to what is said in the dissertation, they retrospectively reason on their study (or remember their thoughts before even reading the paper they used for power analysis - and trust their memories more than the published study) and decide that a plausible effect size is not d = .72, but d = .25. They run a design analysis and conclude that their type M error is very large. I find this course of action very implausible. The researchers, if wise, will probably use d = .72 or some slight downward adjustment of that as plausible effect size (if not wise they will even use d = .90; but that's another story!). In short, if they (or their readers) do not realize that all the effect

sizes that they are encountering on their path (from the published study and from their own data) are heavily inflated, the design analysis will not help them much. If they do realize that all effect sizes in this scenario are inflated, they would be probably very suspicious about their own results with or without retrospective design analysis. Perhaps this type of issue could be discussed a bit more.

Answer: We use this point as an opportunity to further discuss the utility of retrospective design analysis. Retrospective design analysis is useful to evaluate the inferential risks in studies that have already been conducted helping the researchers to interpret the results. Retrospective design analysis should not be confused with post-hoc power analysis. The former defines the plausible effect size according to previous results in the literature or other information external to the study, whereas the latter defines the plausible effect size based on the observed results in the study (a widely-deprecated practice). As underlined by the reviewer, however, the definition of a plausible effect size is very problematic if researchers do not take into account that effect sizes in the literature could be inflated. In Scenario S1 (that is true is a little bit confusing), we wanted to highlight the consequences of using an inflated plausible effect size rather than a more reasonable one. Imagine that researchers plan their sample size considering as plausible effect size $d = .90$ (an inflated value from a single study). In this case, the required sample size to obtain 80% of power it's 20 participants per group, but let's say, without loss of generalizability, they had enough budget to collect 31 participants per group. In this case, the power level is even higher than the canonical 80% (the actual value is 93%), thus researchers would be really confident when interpreting their result. However, imagine that after collecting the data they conducted a retrospective design analysis considering $d = .25$ as plausible effect size (a more reasonable value). In this case, they would obtain much worse results with very low power (16%) and high Type M error (2.6). Now, they will be much more cautious in the interpretation of the results and probably they would not rely on the estimated effect size. Of course, this is a fictional example (why would the researcher use different effect sizes in the power analysis and in the retrospective design analysis?), but we wanted to stress the importance of critically evaluating the literature and carefully reasoning when defining the plausible effect size. However, let's discuss two other examples where retrospective design analysis could be very useful: 1) Retrospective design analysis can be used by researchers when interpreting already published studies to assess the results reliability; 2) In case of complex statistical models, researchers could plan the sample size according to a simplified version of the model (e.g., without considering the covariance structure of the model). After collecting the data, researchers can use retrospective design analysis to assess the reliability of their results considering a range of plausible effect sizes on the fitted model. That is, using the fitted model structure (fixed effects, random effects, and covariance structure of the model estimated on adequately large sample size), the researcher varies only the values of the main effect of interest.

3. Another problem in Scenario 1 is that the description seems to assume that using an effect size from a single study is, by itself, problematic (at p. 24, "researchers should definitely abandon the idea of planning their sample size based on a single published study"). I think that the characteristics of the study are also very

important. For example, an effect found in a single preregistered study with a convincing sampling strategy can be trusted much more than the same effect found in many non-preregistered small-sample studies (or meta-analyses of those studies; Friese & Frankenbach, 2020). I find Scenarios 2 and 3 more convincing because they assume that researchers have some a priori idea that the population effect size is close to d = .25, but fail to collect a sample size as large as desired for other reasons.

Answer: We agree with the reviewer that quantity is not quality. For this reason, we stress again the importance of critically evaluating the literature and carefully reasoning about the definition of the plausible effect size, taking into account all sources of available information. An easy (albeit not really scientific) method to evaluate the quality of published studies is to ask ourselves if we would bet money on the reported effect sizes. I am sure almost everyone would bet on the effect size found in a single preregistered study with a convincing sampling strategy rather than on the effect found in many non-preregistered small-sample studies.

4. At p. 28, the dissertation suggests that power of .51 is problematic (I agree!), but in other parts of the chapter a power of .60 (not very far from that .51!) is presented as reasonable for planning a study (e.g., p. 21).

   Answer: These could indeed appear as inconsistent advice about the optimal level of power required. However, the aim of the chapter is not to provide thresholds but rather to enhance researchers' reasoning about the inferential process. Thus, these examples should not be interpreted as a recommendation. Of course, the higher the merrier, but other aspects should be also considered when defining the required level of power. For example, the cost-value trade-off of the different inferential errors could differ substantially in different scenarios. The same power of 60% could be fine in some cases or absolutely unacceptable in others. The important thing is to always justify the reasons behind the different choices.

3

Design Analysis for Pearson Correlation Coefficient: Evaluating Research Results[1]

Road Map

In the previous chapter, we introduced the Design Analysis, a useful framework that enhances researchers' awareness about the consequence of conducting underpowered studies. In this chapter, we extend the Design Analysis to the case of Pearson's correlation coefficient. This allows us to stress the importance of considering inferential risks related to effect size estimation in a prospective or retrospective design analysis.

3.1 Introduction

Psychological science is increasingly committed to scrutinizing its published findings by promoting large-scale replication efforts, where the protocol of a previous study is repeated as closely as possible with a new sample (Camerer et al., 2016; Camerer et al., 2018; Ebersole et al., 2016; Klein et al., 2014; Klein et al., 2018; Open

[1]This chapter is adapted from Bertoldo et al. (in press), in which I contributed to the development of the original idea, writing of the manuscript, development of the R functions, statistical analysis and the graphical representations. Supplemental Materials available at https://osf.io/b3u8w/. Full reference:
Bertoldo, G., **Zandonella Callegher, C.**, & Altoè, G. (in press). Designing Studies and Evaluating Research Results: Type M and Type S Errors for Pearson Correlation Coefficient [Preprint]. *Meta-Psychology*. https://doi.org/10.31234/osf.io/q9f86

Science Collaboration, 2015). Interestingly, many replication studies found smaller effects than originals (Camerer et al., 2018; Open Science Collaboration, 2015) and among many possible explanations, one relates to a feature of study design: statistical power. In particular, it is plausible for original studies to have lower statistical power than their replications. In the case of underpowered studies, we are usually aware of the lower probability of detecting an effect if this exists, but the less obvious consequences on effect size estimation are often neglected. When underpowered studies are analyzed using thresholds, such as statistical significance levels, effects passing such thresholds have to exaggerate the true effect size (Button et al., 2013; Gelman et al., 2017; Ioannidis, 2008; Ioannidis et al., 2013; Lane & Dunlap, 1978). Indeed, as will be extensively shown below, in underpowered studies only large effects correspond to values that can reject the null hypothesis and be statistically significant. As a consequence, if the original study was underpowered and found an exaggerated estimate of the effect, the replication effect will likely be smaller.

The concept of statistical power finds its natural development in the Neyman-Pearson framework of statistical inference and this is the framework that we adopt in this contribution. Contrary to the Null Hypothesis Significance Testing (NHST), the Neyman-Pearson approach requires to define both the *Null Hypothesis* (i.e., usually, but not necessarily, the absence of an effect) and the *Alternative Hypothesis* (i.e., the magnitude of the expected effect). Further discussion on the Neyman and Pearson approach and a comparison with the NHST is available in Altoè et al. (2020) and Gigerenzer et al. (2004). When conducting hypothesis testing, we usually consider two inferential risks: the Type I error (i.e., the probability α of rejecting the Null Hypothesis if this is true) and the Type II error (i.e., the probability β of not rejecting the Null Hypothesis if this is false). Then, statistical power is defined as the probability 1-β of finding a statistically significant result if the Alternative Hypothesis is true. All this leads to a narrow focus on statistical significance in hypothesis testing, overlooking another important aspect of statistical inference, namely, the effect size estimation.

When effect size estimation is conditioned on the statistical significance (i.e., effect estimates are evaluated only if their p-values are lower than α), effect size exaggeration is a corollary consequence of low statistical power that might not be evident at first. This point can be highlighted considering the Type M (magnitude) and Type S (sign) errors characterizing a study design (Gelman & Carlin, 2014). Given a study design (i.e., sample size, statistical test directionality, α level and plausible effect size formalization), Type M error, also known as *Exaggeration Ratio*, indicates the factor by which a statistically significant effect would be, on average, exaggerated. Type S error indicates the probability to find a statistically significant effect in the opposite direction to the one considered plausible. The analysis that

researchers perform to evaluate the Type M and Type S errors in their research practice is called *design analysis*, given the special focus posed into considering the design of a study (Altoè et al., 2020; Gelman & Carlin, 2014).

Both errors are defined starting from a reasoned guess on the plausible magnitude and direction of the effect under study, which is called *plausible effect size* (Gelman & Carlin, 2014). A plausible effect size is an assumption the researchers make about which is the expected effect in the population. This should not be based on some noisy results from a pilot study but, rather, it could derive from an extensive evaluation of the literature (e.g., theoretical or literature reviews and meta-analyses). When considering the published literature to define the plausible effect size, however, it is important to take into account the presence of publication bias (Franco et al., 2014) and consider techniques for adjusting for the possible inflation of effect size estimates (Anderson et al., 2017). For example if, after taking into account possible inflations, all the main results in a given topic, considering a specific experimental design indicate that the correlation ranges between $r = .15$ and $r = .25$, we could reasonably choose as plausible effect size a value within this range. Or even better, we could consider multiple values to evaluate the results in different scenarios. Note that the definition of the plausible effect size is inevitably highly context-dependent so any attempt to provide reference values would not be useful, instead, it would prevent researchers from reasoning about the phenomenon of interest. Even in extreme cases where no previous information is available, which would question the exploratory/confirmatory nature of the study, researchers could still evaluate which effect would be considered relevant (e.g., from a clinical or economic perspective) and define the plausible effect size according to it.

Why do these errors matter? The concepts of Type M and Type S errors allow enhancing researchers' awareness of a complex process such as statistical inference. Strictly speaking, Design Analysis used in the design phase of a study provides similar information as the classical power analysis, indeed, to a given level of power there is a corresponding Type M and Type S errors. However, it is a valuable conceptual framework that can help researchers to understand the important role of statistical power both when designing a new study or when evaluating previous results from the literature. In particular, it highlights the unwanted (and often overlooked) consequences on effect estimation when filtering for statistical significance in underpowered studies. In these scenarios, there is not only a lower probability of rejecting the null when it is actually false but, even more importantly, any significant result would most likely lead to a misleading overestimation of the actual effect. The exaggeration of effect sizes, in the right or the wrong direction, has important implications on a theoretical and applied level. On a theoretical level, studies' designs with high Type M and Type S errors can foster distorted expectations on the effect under study, triggering a vicious cycle for the planning of future studies.

This point is relevant also for the design of replication studies, which could turn out to be underpowered if they do not take into account possible inflations of the original effect (Button et al., 2013). When studies are used to inform policymaking and real-world interventions, implications can go beyond the academic research community and can impact society at large. In these settings, we could assist to a "hype and disappointment cycle" (Gelman, 2019b), where true effects turn out to be much less impressive than expected. This can produce undesirable consequences on people's lives, a consideration that invites researchers to assume responsibility in effectively communicating the risks related to effects quantification.

To our knowledge, Type M (magnitude) and Type S (sign) errors are not widely known in the psychological research community but their consideration during the research process has the potential to improve current research practices, for example, by increasing the awareness that design choices have on possible studies' results. In a previous work, we illustrated Type M and Type S errors using Cohen's'd as a measure of effect size (Altoè et al., 2020). The purpose of the present contribution is to further increase the familiarity with Type M and Type S errors, considering another common effect size measures in psychology: Pearson correlation coefficient, ρ. We aim to provide an accessible introduction to the Design Analysis framework and enhance the understanding of Type M and Type S errors using several educational examples. The rest of this article is organized as follows: introduction to Type M and Type S errors; description of what is a design analysis and how to conduct one; analysis of Type S and Type M errors when varying alpha levels and hypothesis directionality.

3.2 Type M and Type S Errors

Pearson correlation coefficient is a standardized effect size measure indicating the strength and the direction of the relationship between two continuous variables (Cohen, 1988; Ellis, 2010). Even though the correlation coefficient is widely known, we briefly go over its main features using an example. Imagine that we were interested to measure the relationship between anxiety and depression in a population and we plan a study with n participants, where, for each participant, we measure the level of anxiety (i.e., variable X) and the level of depression (i.e., variable Y). At the end of the study, we will have n pairs of values X and Y. The correlation coefficient helps us answer the questions: how strong is the linear relationship between anxiety and depression in this population? Is the relationship positive or negative? Correlation ranges from -1 to +1, indicating respectively two extreme scenarios of perfect negative relationship and perfect positive relationship[2]. Since the correlation coefficient

[2]Correlation indicates a relationship between variables but does not imply causation. We do not discuss this relevant aspect here but we refer the interested reader to (Rohrer, 2018)

is a dimensionless number, it is a signal to noise ratio where the signal is given by the covariance between the two variables ($cov(x, y)$) and the noise is expressed by the product between the standard deviations of the two variables ($S_x S_y$; see Formula 3.1). In this contribution, following the conventional standards, we will use the symbol ρ to indicate the correlation in the population and the symbol r to indicate the value measured in a sample.

$$r = \frac{cov(x, y)}{S_x \ S_y}. \tag{3.1}$$

Magnitude and sign are two important features characterizing Pearson correlation coefficient and effect size measures in general. And, when estimating effect sizes, errors could be committed exactly regarding these two aspects. Gelman and Carlin (2014) introduced two indexes to quantify these risks:

- Type M error, where M stands for magnitude, is also called Exaggeration Ratio - the factor by which a statistically significant effect is on average exaggerated.
- Type S error, sign - the probability to find a statistically significant result in the opposite direction to the plausible one.

Note that, differently from the other inferential errors, Type M error is not a probability but rather a ratio indicating the average percentage of inflation.

How are these errors computed? In the next paragraphs, we approach this question preferring an intuitive perspective. For a formal definition of these errors, we refer the reader to Altoè et al. (2020), Gelman and Carlin (2014), and Lu et al. (2018). Take as an example the previous fictitious study on the relationship between anxiety and depression and imagine we decide to sample 50 individuals (sample size, n = 50) and to set the α level to 5% and to perform a two-tailed test. Based on theoretical considerations, we expect the plausibly true correlation in the population to be quite strong and positive which we formalize as $\rho = .50$. To evaluate the Type M and Type S errors in this research design, imagine repeating the same study many times with new samples drawn from the same population and, for each study, register the observed correlation (r) and the corresponding p-value.

The first step to compute Type M error is to select only the observed correlation coefficients that are statistically significant in absolute value (for the moment, we do not care about the sign) and to calculate their mean. Type M error is given by the ratio between this mean (i.e., mean of statistically significant correlation coefficients in absolute value) and the plausible effect hypothesized at the beginning, which in this example is $\rho = .50$. Thus, given a study design, Type M error tells us what is the average overestimation of an effect that is statistically significant.

Type S error is computed as the proportion of statistically significant results that have the opposite sign compared to the plausible effect size. In the present example we hypothesized a positive relationship, specifically $\rho = .50$. Then, Type S error is the ratio between the number of times we observed a negative statistically significant result and the total number of statistically significant results. In other words, Type S error indicates the probability to obtain a statistically significant result in the opposite direction to the one hypothesized.

The central and possibly most difficult point in this process is reasoning on what could be the plausible magnitude and direction of the effect of interest. This critical process, which is central also in traditional power analysis, is an opportunity for researchers to aggregate, formalize and incorporate prior information on the phenomenon under investigation (Gelman & Carlin, 2014). What is plausible can be determined on theoretical grounds, using expert knowledge elicitation techniques (see for example O'Hagan, 2019) and consulting literature reviews and meta-analysis, always taking into account the presence of effect sizes inflation in the published literature (Anderson, 2019). Given these premises, it is important to stress that a plausible effect size should not be determined by considering the results of a single study, given the high-level of uncertainty associated with his effect size estimate. The idea is that the plausible effect size should approximate the true effect, which - although never known - can be thought of as "that which would be observed in a hypothetical infinitely large sample" (Gelman & Carlin, 2014, p. 642). For a more exhaustive description of plausible effect size, we refer the interested reader to Altoè et al. (2020) and Gelman and Carlin (2014).

Before we proceed, it is worth noting that there are other recent valuable tools that start from different premises for designing and evaluating studies. Among others, we refer the interested reader to methods which start from the definition of the smallest effect size of interest (SESOI; for a tutorial, see Lakens, Scheel, et al., 2018).

3.3 Design Analysis

Researchers can consider Type M and Type S errors in their practice by performing a *design analysis* (Altoè et al., 2020; Gelman & Carlin, 2014). Ideally, a design analysis should be performed when designing a study. In this phase, it is specifically called *prospective design analysis* and it can be used as a sample size planning strategy where statistical power is considered together with Type M and Type S errors. However, design analysis can also be beneficial to evaluate the inferential risks in studies that have already been conducted and where the study design is known. In these cases, Type M and Type S errors can support results interpretation by communicating the inferential risks in that research design. When design

analysis happens at this later stage, it takes the name of *retrospective design analysis*. Note that retrospective design analysis should not be confused with post-hoc power analysis. A retrospective design analysis defines the plausible effect size according to previous results in the literature or other information external to the study, whereas the post-hoc power analysis defines the plausible effect size based on the observed results in the study and it is a widely-deprecated practice (Gelman, 2019a; S. Goodman & Berlin, 1994).

In the following sections, we illustrate how to perform prospective and retrospective design analysis using some examples. We developed two R functions[3] to perform design analysis for Pearson correlation, which are available at the page https://osf.io/9q5fr/. The function to perform a prospective design analysis is `pro_r()`. It requires as input the plausible effect size (`rho`), the statistical power (`power`), the directionality of the test (`alternative`) which can be set as: "`two.sided`", "`less`" or "`greater`". Type I error rate (`sig_level`) is set as default at 5% and can be changed by the user. The `pro_r()` function returns the necessary sample size to achieve the desired statistical power, Type M error rate, the Type S error probability, and the critical value(s) above which a statistically significant result can be found. The function to perform retrospective design analysis is `retro_r()`. It requires as input the plausible effect size, the sample size used in the study, and the directionality of the test that was performed. Also in this case, Type I error rate is set as default at 5% and can be changed by the user. The function `retro_r()` returns the Type M error rate, the Type S error probability, and the critical value(s)[4].

3.3.1 Case Study

To familiarize the reader with Type M and Type S errors, we start our discussion with a retrospective design analysis of a published study. However, the ideal temporal sequence in the research process would be to perform a prospective design analysis in the planning stage of a research project. This is the time when the design is being laid out and useful improvements can be made to obtain more robust results. In this contribution, the order of presentation aims first, to provide an understanding of how to interpret Type M and Type S errors, and then discuss how they could be taken into account. The following case study was chosen for

[3] An R-package was subsequently developed and now is available on CRAN, PRDA: Conduct a Prospective or Retrospective Design Analysis https://cran.r-project.org/web/packages/PRDA/index.html. PRDA contains other features on Design Analysis, that are beyond the aim of the present paper.

[4] *Critical value* is the name usually employed in hypotheses testing within the Neyman-Pearson framework. In the research practice, this is also known as the *Minimal Statistically Detectable Effect* (Cook et al., 2014; Phillips et al., 2001)

illustrative purposes only and, by no means our objective is to judge the study beyond illustrating an application of how to calculate Type M and Type S errors on a published study.

We consider the study published in *Science* by Eisenberger et al. (2003) entitled: "Does Rejection Hurt? An fMRI Study of Social Exclusion". The research question originated from the observation that the Anterior Cingulate Cortex (ACC) is a region of the brain known to be involved in the experience of physical pain. Could pain from social stimuli, such as social exclusion, share similar neural underpinnings? To test this hypothesis, 13 participants were recruited and each one had to play a virtual game with other two players while undergoing functional Magnetic Resonance Imaging (fMRI). The other two players were fictitious, and participants were actually playing against a computer program. Players had to toss a virtual ball among each other in three conditions: social inclusion, explicit social exclusion and implicit social exclusion. In the social inclusion condition, the participant regularly received the ball. In the explicit social exclusion condition the participant was told that, due to technical problems, he was not going to play that round. In the implicit social exclusion condition, the participant experienced being intentionally left out from the game by the other two players. At the end of the experiment, each participant completed a self-report measure regarding their perceived distress when they were intentionally left out by the other players. Considering only the implicit social exclusion condition, a correlation coefficient was estimated between the measure of distress and neural activity in the Anterior Cingulate Cortex. As suggested by the large and statistically significant correlation coefficient between perceived distress and activity in the ACC, r = .88, p ¡ .005 (Eisenberger et al., 2003, p. 291), authors concluded that social and physical pain seem to share similar neural underpinnings.

Before proceeding to the retrospective design analysis, we refer the interested reader to some background history regarding this study. This was one of the many studies included in the famous paper "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (Vul et al., 2009) which raised important issues regarding the analysis of neuroscientific data. In particular, this paper noted that the magnitude of correlation coefficients between fMRI measures and behavioural measures were beyond what could be considered plausible. We refer the interested reader also to the commentary by Yarkoni (2009), who noted that the implausibly high correlations in fMRI studies could be largely explained by the low statistical power of experiments.

A retrospective design analysis should start with thorough reasoning on the plausible size and direction of the effect under study. To produce valid inferences, a lot of attention should be devoted to this point by integrating external information. For the sake of this example, we turn to the considerations made by Vul and Pashler

(2017) who suggested correlations between personality measures and neural activity to be likely around $\rho = .25$. A correlation of $\rho = .50$ was deemed plausible but optimistic and a correlation of $\rho = .75$ was considered theoretically plausible but unrealistic.

3.3.2 Retrospective Design Analysis

To perform a retrospective design analysis on the case study, we need information on the research design and the plausible effect size. Based on the previous considerations, we set the plausible effect size to be $\rho = .25$. Information on the sample size was not available in the original study (Eisenberger et al., 2003) and was retrieved from Vul et al. (2009) to be $n = 13$. The α level and the directionality of the test were not reported in the original study, so for the purpose of this example, we will consider $\alpha = .05$ and a two-tailed test. Given this study design, what are the inferential risks in terms of effect size estimation?

We can use the R function `retro_r()`, whose inputs and outputs are displayed below[5]. In this study, the statistical power is .13, that is to say, there is a 13% probability to reject the null hypothesis, if an effect of at least $\rho = |.25|$ exists. Consider this point together with the results obtained in the experiment: $r = .88$, $p < .005$ (Eisenberger et al., 2003, p. 291). It is clear that, even though the probability to reject the null hypothesis is low (power of 13%), this event could happen. And when it does happen, it is tempting to believe that results are even more remarkable (Gelman & Loken, 2014). However, this design comes with serious inferential risks for the estimation of effect sizes, which could be grasped by presenting Type M and Type S errors. A glance at their value communicates that it is not impossible to find a statistically significant result, but when it does happen, the effect sizes could be largely overestimated - Type M = 2.58 - and maybe even in the wrong direction - Type S = .03. The Type M error rate of 2.58 indicates that a statistically significant correlation is on average about two and a half times the plausible value. In other words, statistically significant results emerging in such a research design will on average overestimate the plausible correlation coefficient by 160%. The Type S error of .03 suggests that there is a three percent probability to find a statistically significant result in the opposite direction, in this example, a negative relationship.

```
retro_r(rho = .25, n = 13, alternative = "two.sided",
        sig_level = .05, seed = 2020)

##
##  Design Analysis
```

[5]The option `seed` allows setting the random number generator to obtain reproducible results.

```
##
## Hypothesized effect: rho = 0.25
##
## Study characteristics:
##    n    alternative   sig_level
##    13   two.sided     0.05
##
## Inferential risks:
##    power   typeM   typeS
##    0.127   2.583   0.028
##
## Critical value(s): r =  ±0.553
```

In this research design, the critical values above which a statistically significant result is declared correspond to $r = \pm.55$. These values are highlighted in Figure 3.1 as the vertical lines in the sampling distribution of correlation coefficients under the null hypothesis. Notice that the plausible effect size lies in the region of acceptance of the null hypothesis[6]. Therefore, it is impossible to simultaneously find a statistically significant result and estimate an effect close to the plausible one ($\rho = .25$). The figure represents the so-called Winner's curse: "the 'lucky' scientist who makes a discovery is cursed by finding an inflated estimate of that effect" (Button et al., 2013).

3.3.3 Prospective Design Analysis

Ideally, Type M and Type S errors should be considered in the design phase of a study during the decision-making process regarding the experimental protocol. At this stage, prospective design analysis can be used as a sample size planning strategy which aims to minimize Type M and Type S errors in the upcoming study.

Imagine that we were part of the research team in the previous case study exploring the relationship between activity in the Anterior Cerebral Cortex and perceived distress. When drafting the research protocol, we face the inevitable discussion on how many participants we are going to recruit. This choice depends on available resources, type of study design, constraints of various nature and, importantly, the plausible magnitude and direction of the phenomenon that we are going to study. As previously mentioned, deciding on a plausible effect size is a fundamental step which requires great effort and should not be done by trying different values only to obtain a more desirable sample size. Instead, proposing a

[6]Note that accepting the null hypothesis is possible only in the Neyman-Pearson approach and not in the NHST

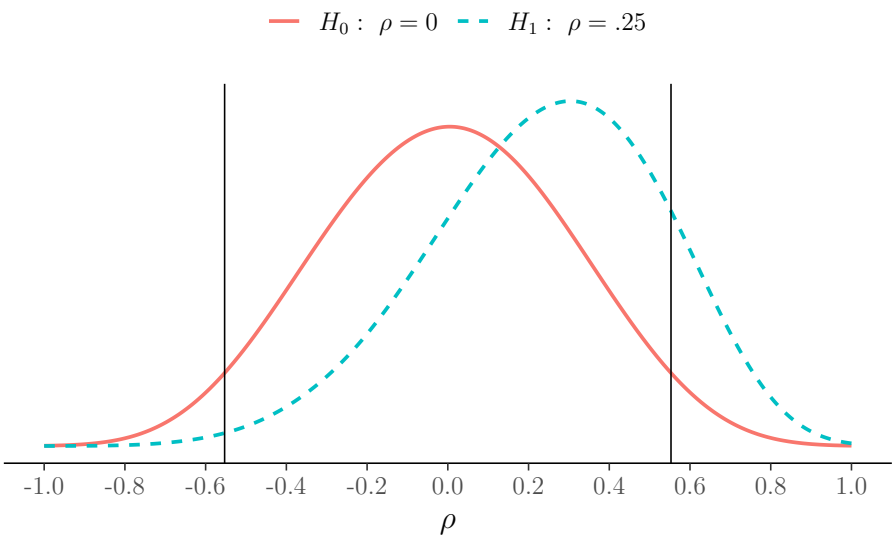

Figure 3.1: Winner's course. H_0 = Null Hypothesis, H_1 = Alternative Hypothesis. When sample size, directionality of the test and Type I error probability are set, also the smallest effect size above which is possible to find a statistically significant result is set. In this case, the plausible effect size, $\rho = .25$, lies in the region where it is not possible to reject H_0 (the region delimited by the two vertical lines). Thus, it is impossible to simultaneously find a statistically significant result and an effect close to the plausible one. In other words, a statistically significant effect must exaggerate the plausible effect size.

plausible effect size is where the expert knowledge of the researcher can be formalized and can greatly contribute to the informativeness of the study that is being planned. For the sake of these examples, we adopt the previous consideration and we suppose that common agreement is reached on a plausible correlation coefficient to be around $\rho = .25$. Finally, we would like to leave open the possibility to explore whether the relationship goes in the opposite direction to the one hypothesized, so we decide to perform a two-tailed test.

We can implement the prospective design analysis using the function `pro_r()` which inputs and outputs are displayed below. About 125 participants are necessary to have 80% probability to detect an effect of at least $\rho = \pm.25$ if it actually exists. With this sample size, the Type S error is minimized and approximates zero. In this study design, the Type M error is 1.11 indicating that statistically significant results are on average exaggerated by 11%. It is possible to notice that the critical values are $r = \pm.18$, further highlighting that our plausible effect size is actually included among those values that lead to the acceptance of the alternative hypothesis.

```
pro_r(rho = .25, power = .8, alternative = "two.sided",
      sig_level = .05, seed = 2020)

##
##  Design Analysis
##
## Hypothesized effect: rho = 0.25
##
## Study characteristics:
##    n     alternative   sig_level
##    125   two.sided     0.05
##
## Inferential risks:
##    power   typeM   typeS
##    0.806   1.111   0
##
## Critical value(s): r =  ±0.176
```

In a design analysis, it is advisable to investigate how the inferential risks would change according to different scenarios in terms of statistical power and plausible effect size. Changes in both these factors impact Type M and Type S errors. For example, maintaining the plausible correlation of $\rho = .25$, if we decrease statistical power from .80 to .60 only 76 participants are required (see Table 3.1). However, this is associated with an increased Type M error rate from 1.11 to 1.28. That is to say, with 76 subjects the plausible effect size will be on average overestimated by 28%. Alternatively, imagine that we would like to maintain a statistical power of 80%, what happens if the plausible effect size is slightly larger or smaller? The necessary sample size would spike to 344 for a $\rho = .15$ and decrease to 60 for $\rho = .35$. In both scenarios, the Type M error remains about 1.12, which reflects the more general point that for 80% power, Type M error is around 1.10. In all these scenarios, Type S error is close to zero, hence not worrisome.

For completeness, Figure 3.2 summarizes the relationship between statistical power, Type M and Type S errors as a function of sample size in three scenarios of plausible correlation coefficients. We display the three values that Vul and Pashler (2017) considered for correlations between fMRI measures and behavioural measures with different degrees of plausibility. An effect of $\rho = .75$ was deemed theoretically plausible but unrealistic, $\rho = .50$ was more plausible but optimistic, and $\rho = .25$ was more likely. The curves illustrate a general point: Type M and Type S error increase with smaller sample sizes, smaller plausible effect sizes and lower statistical power. Also, the figure shows that statistical power, Type M and Type S errors are related to each other: as power increases, Type M and Type S errors decrease.

Table 3.1: Prospective design analysis in different scenarios of plausible effect size and statistical power.

ρ	Power	Sample Size	Type M	Type S	Critical r value
0.25	0.6	76	1.280	0	±0.226
0.15	0.8	344	1.116	0	±0.106
0.35	0.8	60	1.115	0	±0.254

Note: In all cases, alternative = `"two.sided"` and sig_level = .05.

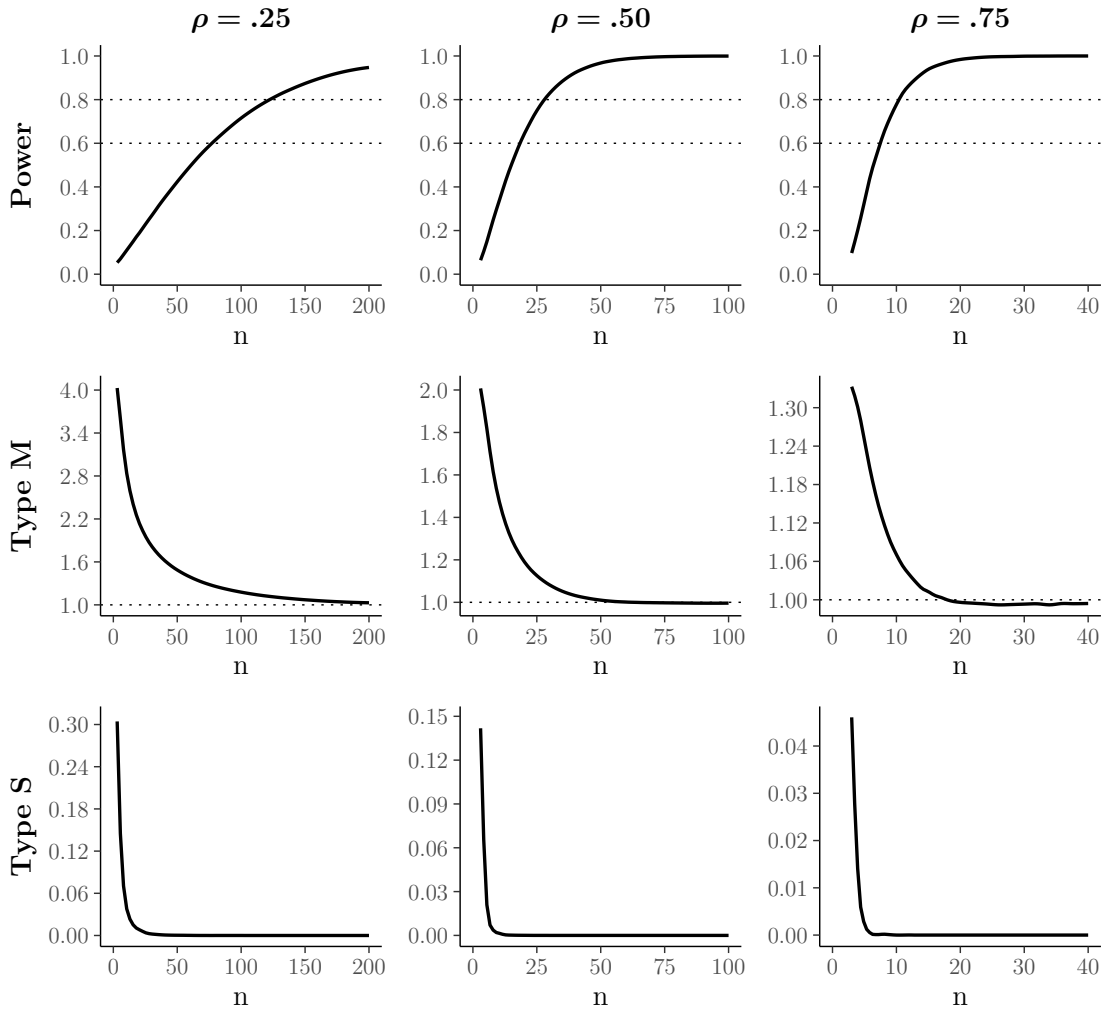

Figure 3.2: How Type M, Type S and Statistical power vary as a function of sample size in three different scenarios of plausible effect size (ρ = .25, ρ = .50, ρ = .75). Note that, for the sake of interpretability, we decided to use different scales for both the x-axis and y-axis in the three scenarios of plausible effect size.

At first, it might seem that Type M and Type S errors are redundant with the information provided by statistical power. Even though they are related, we believe that Type M and Type S errors bring added value during the design phase of a research protocol because they facilitate a connection between how a study is planned and how results will actually be evaluated. That is to say, final results will comprise of a test statistics with an associated p-value and effect size measure. If the interest is maximizing the accuracy with which effects will be estimated, then Type M and Type S errors directly communicate the consequences of design choices on effect size estimation.

3.4 Varying α levels and Hypotheses Directionality

So far, we did not discuss two other important decisions that researchers have to take when designing a study: statistical significance threshold or α level, and directionality of the statistical test, one-tailed or two-tailed. In this section, we illustrate how different choices regarding these aspects impact Type M and Type S errors.

A lot has been written regarding the automatic adoption of a conventional α level of 5% (e.g., Gigerenzer et al., 2004; Lakens, Adolfi, et al., 2018). This practice is increasingly discouraged, and researchers are invited to think about the best trade-off between α level and statistical power, considering the aim of the study and available resources. The α level impacts Type M and Type S errors as much as it impacts statistical power. Everything else equal, Type M error increases with decreasing α level (i.e., negative relationship), whereas Type S error decreases with decreasing α level (i.e., positive relationship). To further illustrate the relation between Type M error and α level, let us take as an example the previous case study with a sample of 13 participants, plausible effect size $\rho = .25$ and two-tailed test. Table 3.2 shows that by lowering the α level from 10% to .10%, the critical values move from $r = \pm.48$ to $r = \pm.80$. This suggests that, with these new higher thresholds, the exaggeration of effects will be even more pronounced because effects have to be even larger to pass such higher critical values (i.e., higher Type M error). Instead, the relationship between Type S error and α level can be clarified thinking that by lowering the statistical significance threshold, we are being more conservative to falsely reject the null hypothesis in general which implies that we are also being more conservative to falsely reject the null hypothesis in the wrong direction.

Another important choice in study design is the directionality of the test (i.e., one-tailed or two-tailed). Design analysis invites reasoning on the plausible effect

Table 3.2: How changes in α level impact Power, Type M error, Type S error and critical values.

α-level	Power	Type M	Type S	Critical r value
0.100	0.212	2.369	0.040	±0.476
0.050	0.127	2.583	0.028	±0.553
0.010	0.035	2.977	0.011	±0.684
0.005	0.021	3.088	0.014	±0.726
0.001	0.005	3.340	0.000	±0.801

Note: In all cases, $\rho = .25$, n = 13, and alternative = `"two.sided"`.

size and hypothesizing the direction of the effect, not only its magnitude. So why should a researcher perform non-directional statistical tests when there is a hypothesized direction? Performing a two-tailed test leaves open the possibility to find an unexpected result in the opposite direction (Cohen, 1988), a possibility which may be of special interest for preliminary exploratory studies. However, in more advanced stages of a research program (i.e., confirmatory study), directional hypotheses benefit from higher statistical power and lower Type M error rates (Figure 3.3). As an example, let us consider the differences between a two-tailed test and a one-tailed test in the previous case study. We can perform a new prospective design analysis (see code below) with a plausible correlation of $\rho = .25$, 80% statistical power, but this time setting the argument `alternative` in the R function to `\greater"`. A comparison with the previous prospective design analyses, suggests that the same Type M error rate of about 10% is guaranteed with 94 participants, instead of the 125 subjects necessary with a two-tailed test. Note that Type S error is not possible in directional statistical tests. Indeed, all the statistically significant results are obtainable only in the hypothesized direction, not the opposite one.

```
pro_r(rho = .25, power = .8, alternative = "greater",
      sig_level = .05, seed = 2020)

##
##  Design Analysis
##
## Hypothesized effect: rho = 0.25
##
## Study characteristics:
##    n    alternative   sig_level
##    94   greater       0.05
##
```

```
## Inferential risks:
##    power   typeM   typeS
##    0.793   1.14    0
##
## Critical value(s): r =  0.171
```

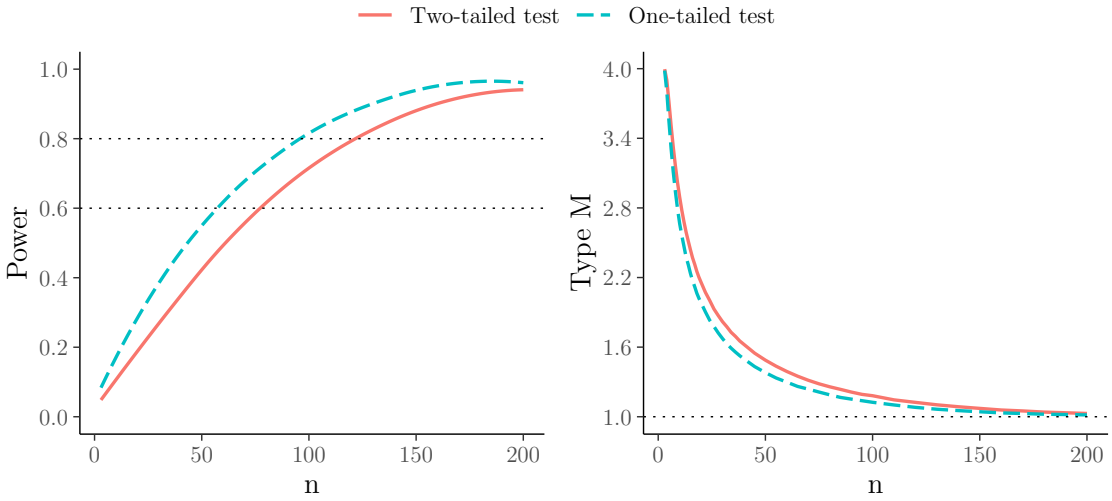

Figure 3.3: Comparison of Type M error rate and Power level between one-tailed and two-tailed test with $\rho = .25$, $\alpha = .05$. n = sample size.

Valid conclusions require decisions on test directionality and α level to be taken a priori, not while data are being analyzed (Cohen, 1988). These decisions can take place during a prospective design analysis, which aligns with the increasing interest in psychological science to transparently communicate and justify design choices through studies' preregistration in public repositories (e.g., Open Science Framework; Aspredicted.com). Preregistration of studies' protocol is particularly valuable for researchers endorsing an error statistics philosophy of science, where the evaluation of research results takes into account the severity with which claims are tested (Lakens, 2019; Mayo, 2018). Severity depends on the degree to which a research protocol tries to falsify a claim. For example, a one-tailed statistical test provides greater severity than a two-tailed statistical test. As noted by Lakens (2019), preregistration is important to openly share a priori decisions, such as test-directionality, providing valuable information for researchers interested in evaluating the severity of research claims.

3.5 Publication Bias and Significance Filter

On a concluding note, we would like to clarify the relationship of Design Analysis with publication bias and the statistical significance filter.

While publication bias and Type M and Type S errors are related, they operate at two different levels. Publication bias refers to a publication system that favours statistically significant results over non-statistically significant findings. This phenomenon alone cannot explain the presence of exaggerated effects. Imagine if all studies in the literature were conducted with high statistical power, then statistically significant findings would probably not be so extreme. The problem of exaggerated effect sizes in the literature can be explained only by a combination of publication bias with studies that have low statistical power. As previously shown, statistical power and Type M and Type S errors are related to each other: low statistical power corresponds to higher Type M and Type S errors.

The critical element is the application of the statistical significance filter without taking into account statistical power. Design Analysis per se does not solve this issue but, instead, it allows us to recognize its problematic consequences. In the same way as statistical power is a characteristic of a study design, so are Type M and Type S errors, however, the two are qualitatively different in terms of the kind of reasoning they favour. Statistical power is defined in terms of probability of rejecting the Null hypothesis and, even though this is based on an effect size of interest, the relationship "low power - high possibility of exaggeration" may not be straightforward for everyone. Instead, Type M and Type S errors directly quantify the possible exaggeration. Furthermore, their consideration protects against another possible pitfall. When in a study a statistically significant result is found and the associated effect size estimate is large, the finding could be interpreted as robust and impressive. However, this interpretation is not always appropriate. Here, the missing piece of information is statistical power. If power is considered, researchers would realize that a large effect was found in a context where there was a low probability to find it. But this interpretation is not explicitly stating an important aspect: in these conditions, the only way to find a statistically significant result is by overestimating the true effect. On the contrary, this consequence becomes immediately clear once Type M and Type S errors are considered retrospectively. Similarly, considering Type M and Type S prospectively favours reasoning in terms of effect size rather than the probability of rejecting the null hypothesis when setting the sample size in a design analysis.

3.6 Discussion and Conclusion

In the scientific community, it is quite widespread the idea that the literature is affected by a problem with effect size exaggeration. This issue is usually explained in terms of studies' low statistical power combined with the use of thresholds of statistical significance (Button et al., 2013; Ioannidis, 2008; Ioannidis et al., 2013; Lane & Dunlap, 1978; Yarkoni, 2009; Young et al., 2008). Statistically significant results can be obtained even in underpowered studies and it is precisely in these cases that we should worry the most about issues of overestimation. Type M and Type S errors quantify and highlight the inferential risks directly in terms of effect size estimation, which are implied by the concept of statistical power but might not be recognizable outright. So far, only a handful of papers explicitly mentioned Type M and Type S errors (Altoè et al., 2020; Gelman, 2018; Gelman & Carlin, 2013, 2014; Gelman et al., 2017; Gelman & Tuerlinckx, 2000; Lu et al., 2018; Vasishth et al., 2018). With the broader goal of facilitating their consideration in psychological science, in the present contribution we illustrated how Type M and Type S errors are considered in a design analysis using one of the most common effect size measures in psychology, Pearson correlation coefficient.

Peculiar to design analysis is the focus on the implications of design choices on effect sizes estimation rather than statistical significance only. We illustrated how Type M and Type S errors can be taken into account with a *prospective design analysis*. In the planning stage of a research project, design analysis has the potential to increase researchers' awareness of the consequences that their sample size choices have on uncertainty about final estimates of the effects. This favours reasoning in similar terms to those in which results will be evaluated, that is to say, effect size estimation. But understanding the inferential risks in a study design is also beneficial once results are obtained. We presented *retrospective design analysis* on a published study, and the same process can be useful for studies in general, especially those ending without the necessary sample size to maximize statistical power and minimize Type M and Type S errors. In all cases, presenting their values effectively communicates the uncertainty of the results. In particular, Type M and Type S errors put a red flag when results are statistically significant, but the effect size could be largely overestimated and in the wrong direction. Finally, both prospective and retrospective design analysis favours cumulative science encouraging the incorporation of expert knowledge in the definition of the plausible effect sizes.

It is important to remark that even if Design Analysis is based on the definition of a plausible effect size, a best practice should be to conduct multiple Design Analyses by considering different scenarios which include different plausible effect sizes and levels of power to maximize the informativeness of both a prospective and a retrospective analysis.

To make design analysis accessible to the research community, we provide the R functions to perform prospective design analysis and retrospective design analysis for Pearson correlation coefficient https://osf.io/9q5fr/.

Finally, prospective design analysis could contribute to better research design, however many other important factors were not considered in this contribution. For example, the validity and reliability of measurements should be at the forefront in research design, and careful planning of the entire research protocol is of utmost importance. Future works could tackle some of these shortcomings for example, including an analysis of the quality of measurement on the estimates of Type M and Type S errors. Also, we believe that it would be valuable to provide extension of design analysis for other common effect size measures with the development of statistical software packages that could be directly used by researchers. Moreover, design analysis on Pearson correlation can be easily extended to the multivariate case where multiple predictors are considered. Lastly, design analysis is not limited to the Neyman-Pearson framework but can be considered also within other statistical approaches such as Bayesian approach. Future works could implement design analysis to evaluate the inferential risks related to the use of Bayes Factors and Bayesian Credibility Intervals.

Summarizing, choices regarding studies' design impact effect size estimation and Type M (magnitude) error and Type S (sign) error allow to directly quantify these inferential risks. Their consideration in a prospective design analysis increases awareness of what are the consequences of sample size choice reasoning in similar terms to those used in results evaluation. Instead, retrospective design analysis provides further guidance on interpreting research results. More broadly, design analysis reminds researchers that statistical inference should start before data collection and does not end when results are obtained.

4

PRDA: An R package for Prospective and Retrospective Design Analysis[1]

Road Map

In the previous chapters, we introduced the design analysis framework. In this chapter, we present the `PRDA` R-package. This package allows performing prospective and retrospective design analysis in the case of Pearson's correlation between two variables or mean comparisons.

4.1 Introduction

Design Analysis was introduced by Gelman and Carlin (2014) as an extension of Power Analysis. Traditional power analysis has a narrow focus on statistical significance. Design analysis, instead, evaluates together with power levels also other inferential risks (i.e., Type M error and Type S error), to assess estimates uncertainty under hypothetical replications of a study.

[1]This chapter is adapted from Zandonella Callegher et al. (2021), in which I contributed to the development of the original idea, writing of the manuscript, development of the R functions, statistical analysis and the graphical representations. GitHub repository https://github.com/ClaudioZandonella/PRDA. Full reference:
Zandonella Callegher, C., Bertoldo, G., Toffalini, E., Vesely, A., Andreella, A., Pastore, M., & Altoè, G. (2021). PRDA: An R package for Prospective and Retrospective Design Analysis. *Journal of Open Source Software*, 6(58), 2810. https://doi.org/10.21105/joss.02810

Given an hypothetical value of effect size and study characteristics (i.e., sample size, statistical test directionality, significance level), *Type M error* (Magnitude, also known as Exaggeration Ratio) indicates the factor by which a statistically significant effect is on average exaggerated. *Type S error* (Sign), instead, indicates the probability of finding a statistically significant result in the opposite direction to the hypothetical effect.

Although Type M error and Type S error depend directly on power level, they underline valuable information regarding estimates uncertainty that would otherwise be overlooked. This enhances researchers awareness about the inferential risks related to their studies and helps them in the interpretation of their results. However, design analysis is rarely applied in real research settings also for the lack of dedicated software.

To know more about design analysis consider Gelman and Carlin (2014) and Lu et al. (2018). While, for an introduction to design analysis with examples in psychology see Altoè et al. (2020) and Bertoldo et al. (in press).

4.2 Statement of Need

`PRDA` is an R package performing prospective or retrospective design analysis to evaluate inferential risks (i.e., power, Type M error, and Type S error) in a study considering Pearson's correlation between two variables or mean comparisons (one-sample, paired, two-sample, and Welch's t-test). *Prospective Design Analysis* is performed in the planning stage of a study to define the required sample size to obtain a given level of power. *Retrospective Design Analysis*, instead, is performed when the data have already been collected to evaluate the inferential risks associated with the study.

Another recent R package, `retrodesign` (Timm, 2019), allows conducting retrospective design analysis considering estimate of the unstandardized effect size (i.e., regression coefficient or mean difference) and standard error of the estimate. `PRDA` package, instead, considers standardized effect size (i.e., Pearson correlation coefficient or Cohen's d) and study sample size. These are more commonly used in research fields such as Psychology or Social Science, and therefore are implemented in `PRDA` to facilitate researchers' reasoning about design analysis. `PRDA`, additionally, offers the possibility to conduct a prospective design analysis and to account for the uncertainty about the hypothetical value of effect size. In fact, hypothetical effect size can be defined as a single value according to previous results in the literature or experts indications, or by specifying a distribution of plausible values.

The package is available from GitHub (https://github.com/ClaudioZandonella/PRDA) and `CRAN` (https://CRAN.R-project.org/package=PRDA). Documentation about the package is available at https://claudiozandonella.github.io/PRDA/.

4.3 Examples

Imagine a study evaluating the relation a given personality trait (e.g., introversion) and math performance. Suppose that 20 participants were included in the study and results indicated a statistically significant correlation (e.g, $r = .55, p = .012$). The magnitude of the estimated correlation, however, is beyond what could be considered plausible in this field.

4.3.1 Retrospective Design Analysis

Suppose previous results in the literature indicate correlations in this area are more likely to be around $\rho = .25$. To evaluate the inferential risks associated with the study design, we can use the function `retrospective()`.

```
library(PRDA)

set.seed(2020) # set seed to make results reproducible

retrospective(effect_size = .25, sample_n1 = 20,
              test_method = "pearson")

##
##  Design Analysis
##
## Hypothesized effect:  rho = 0.25
##
## Study characteristics:
##    test_method   sample_n1   sample_n2   alternative   sig_level   df
##    pearson       20          NULL        two_sided     0.05        18
##
## Inferential risks:
##    power   typeM   typeS
##    0.185   2.161   0.008
##
## Critical value(s): rho  =  ± 0.444
```

In the output, we have the summary information about the hypothesized population effect, the study characteristics, and the inferential risks. We obtained a statistical power of almost 20% that is associated with a Type M error of around 2.2 and a Type S error of 0.01. That means, statistical significant results are on average an overestimation of 120% of the hypothesized population effect and there is a 1% probability of obtaining a statistically significant result in the opposite di-

rection. To know more about function arguments and examples see the function documentation and vignette.

Effect Size Distribution

Alternatively, if no precise information about hypothetical effect size is available, researchers could specify a distribution of values to account for their uncertainty. For example, they might define a normal distribution with mean of .25 and standard deviation of .1, truncated between .10 and 40.

```
retrospective(effect_size = function(n) rnorm(n, .25, .1),
              sample_n1 = 20, test_method = "pearson",
              tl = .1, tu = .4, B = 1e3,
              display_message = FALSE)

## Truncation could require long computational time

##
##  Design Analysis
##
## Hypothesized effect:  rho ~ rnorm(n, 0.25, 0.1) [tl =  0.1 ; tu = 0.4 ]
##    n_effect   Min.    1st Qu.   Median   Mean    3rd Qu.   Max.
##    1000       0.101   0.197     0.25     0.252   0.308     0.4
##
## Study characteristics:
##    test_method   sample_n1   sample_n2   alternative   sig_level   df
##    pearson       20          NULL        two_sided     0.05        18
##
## Inferential risks:
##         Min.    1st Qu.   Median   Mean       3rd Qu.   Max.
## power   0.055   0.133     0.1880   0.203727   0.26600   0.449
## typeM   1.407   1.785     2.1645   2.347745   2.70075   5.263
## typeS   0.000   0.000     0.0060   0.017573   0.02300   0.246
##
## Critical value(s): rho  =  ± 0.444
```

Consequently this time we obtained a distribution of values for power, Type M error, and Type S error. Summary information are provided in the output and sampled effects and corresponding error values are available in the returned object.

In `PRDA` there are no implemented functions to obtain graphical representations of the results. However, it is easy to access all the results and use them to create the plots according to your own needs and preferences. Plotting the distribution of sampled effects, researchers can evaluate whether they accurately represent the

intended distribution (see Figure 4.1). If not, the number of sampled effects should be increased (`B_effect`).

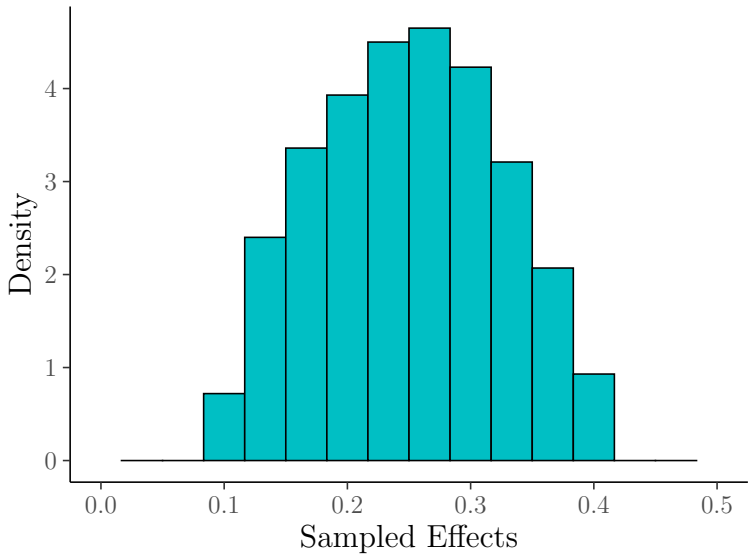

Figure 4.1: Distribution of the sampled effects according to $\mathcal{N}(.25, .1)$ truncated between .10 and .40.

Also the values of Power, Type M error, and Type S error can be plotted to evaluate their distributions (see Figure 4.2).

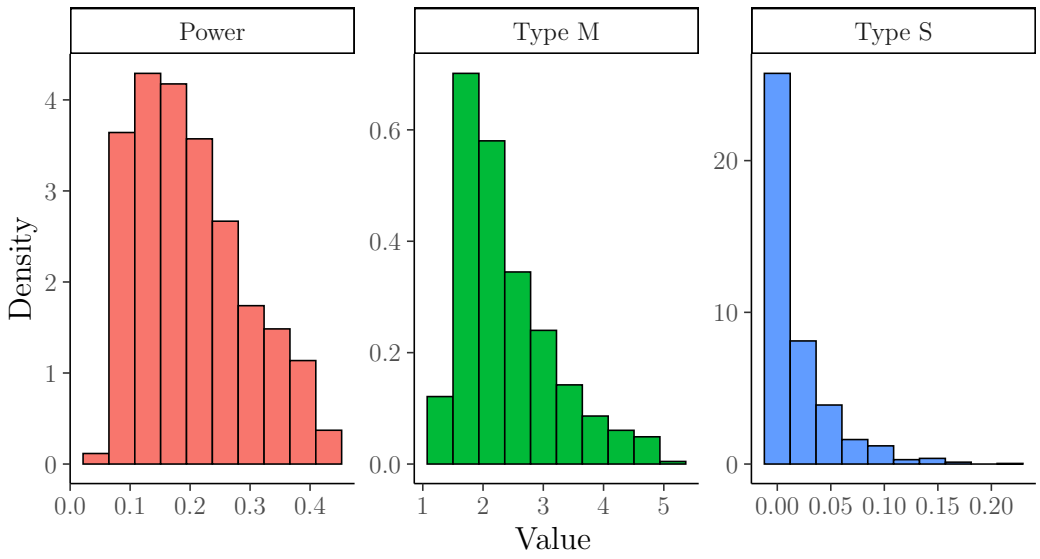

Figure 4.2: Distribution of Power, Type M error, and Type S error.

4.3.2 Prospective Design Analysis

Given the previous results, researchers might consider planning a replication study to obtain more reliable results. The function `prospective()` can be used to compute the sample size needed to obtain a given level of power (e.g., power = 80%).

```
prospective(effect_size = .25, power = .8,
            test_method = "pearson", display_message = FALSE)

##
##  Design Analysis
##
## Hypothesized effect:  rho = 0.25
##
## Study characteristics:
##    test_method   sample_n1   sample_n2   alternative   sig_level   df
##    pearson       122         NULL        two_sided     0.05        120
##
## Inferential risks:
##    power   typeM   typeS
##    0.796   1.12    0
##
## Critical value(s): rho  =  ± 0.178
```

In the output, we have again the summary information about the hypothesized population effect, the study characteristics, and the inferential risks. To obtain a power of around 80% the required sample size is $n = 122$, the associated Type M error is around 1.10 and the Type S error is approximately 0. To know more about function arguments and examples see the function documentation and vignette.

4.4 Conclusions

The design analysis framework is a useful conceptual tool to enhance researchers' awareness about the consequence of conducting underpowered studies. In particular, what could pass unnoticed is that, in case of underpowered studies, there is not only a higher probability of not rejecting the Null Hypothesis if this is false, but, even more importantly, there is also a higher risk of obtaining misleading estimates in case of significant results. Type M error (and Type S error) allows us to directly highlight this important issue. `PRDA` allows users to get familiar with the concepts of design analysis considering common cases as the evaluation of Pearson's correlation between two variables or mean comparisons.

Part II

Model Comparison

5

Model Comparison via Information Criteria: A Different Approach to Hypothesis Testing[1]

Road Map

In the previous chapters, we introduced the Design Analysis to enhance researchers' awareness about the inferential process in the NHST and the consequence on effect size estimation when conducting underpowered studies. In this chapter, we further discuss the limits of the NHST approach in the evaluation of research hypotheses. To overcome these issues, we introduce the model comparison approach using the information criteria that allows to properly evaluate the research hypotheses. As a case study, we consider the stereotype threat effects on Italian girls' mathematics performance.

[1]This chapter is adapted from of Agnoli et al. (2021) and its supplemental material available at https://osf.io/3u2jd/, in which I contributed to the writing of the supplemental material, the statistical analyses and the graphical representations. In particular, a special emphasis is dedicated to the statistical approach used. Full reference:
Agnoli, F., Melchiorre, F., **Zandonella Callegher, C.**, & Altoè, G. (2021). Stereotype threat effects on Italian girls' mathematics performance: A failure to replicate. *Developmental Psychology*, 57(6), 940–950. https://doi.org/10.1037/dev0001186

5.1 Introduction

The Null Hypothesis Significance Testing (NHST) is the dominant statistical approach in Social and Psychological sciences (Chavalarias et al., 2016). In the literature, however, many problematic aspects of the NHST have been highlighted (Szucs & Ioannidis, 2017; Wasserstein et al., 2019). In particular, the misunderstanding and abuse of the NHST, in what Gigerenzer defined *"the null ritual"* 2004, is considered to be one of the main causes behind the replication crisis in social and psychological sciences.

The NHST places a narrow focus on statistical testing leading to dichotomous thinking significant/non-significant. Most of the researchers, however, do not interpret correctly the meaning of the p-values and the implications of statistically significant results more in general (Gigerenzer et al., 2004). The p-value does not quantify the probability of one hypothesis given the data, i.e. $P(H|D)$, but the probability of the data given one hypothesis, i.e. $P(H|D)$. This often leads to misleading interpretations and false beliefs about the validity of the results. In fact, researchers are usually interested in evaluating their hypotheses of interest but the p-value does not quantify the evidence in favour of one hypothesis but only against it. Thus, given a statistically significant result, they usually interpret it as evidence in favour of their hypotheses. In reality, however, nothing can be said about the plausibility of their hypotheses. This is one of the main limits of the NHST as it does not allow to answer the question the researchers are more interested in, that is the evaluation of research or theoretical hypotheses.

Moreover, researchers might easily underestimate the occurrence of false-positive results. Statistically significant results may occur even just by chances for many different factors other than the presence of a real effect. The researcher degrees of freedom, Questionable Research Practices (QRPs; John et al., 2012), questionable measurement practices (Flake & Fried, 2020; Schimmack, 2021), and misuse of statistical techniques, all can contribute to statistically significant results without an actual effect being present.

Even in the case of a true effect, filtering for statistical significance may lead to unreliable results. In fact, in the case of underpowered studies, statistically significant results are almost surely an overestimation of the actual effect. This aspect is usually neglected in the NHST approach and also in the traditional power analysis, given their narrow focus statistical significance. To highlight this problem, Gelman and Carlin (2014) introduced the *Design Analysis*. The design analysis enhances researchers' awareness about the consequence of conducting underpowered studies evaluating the inferential risks related to effect size estimation (Altoè et al., 2020; Bertoldo et al., in press). The Design Analysis per se, however, does not solve the issues related to the NHST, but it only helps to highlight its consequences. To overcome NHST limits, therefore, we need to move away from significance test-

ing towards the evaluation of informative hypotheses using the model comparison approach.

5.1.1 Model Comparison Approach

In the model comparison approach, first, the research hypotheses must be formalized as statistical models. According to the predictors included in the model, researchers define the variables expected to have an important role in the phenomenon of interest. This is an important step as it forces researchers to define appropriate statistical models that reflect the data generative process specifying and clarifying all the underlying assumptions of the hypotheses of interest.

Subsequently, the obtained models are compared in terms of the statistical evidence (i.e., support by the data) using for example the information criteria (Wagenmakers & Farrell, 2004). Information criteria provide an estimate of the average deviance (i.e., error) of a model's ability to predict new data, and thus lower values are interpreted as evidence of a better model (McElreath, 2020b). This allows us to consider the trade-off between parsimony and goodness-of-fit (Vandekerckhove et al., 2015) when evaluating models; as the complexity of a model increases (i.e., more parameters), its fit to the data increases, but generalizability (i.e., ability to predict new data) decreases. In statistics, this issue is often referred to as the *trade-off between bias and variaince*, where bias is related to underfitting and variance is related to overfitting the data (Azzalini et al., 2012; McElreath, 2020b). The researchers aim to find the right balance between fit and generalizability to describe, with a statistical model, the important features of the studied phenomenon, but not the random noise of the observed data. Model comparison favours models with effects that offer an appropriate description of the data generating process, penalizing the inclusion of further, unnecessary effects that only introduce an unnecessary level of complexity.

Commonly used information criteria are, for example, the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). These criteria can be used to select the most plausible models among the considered models, given the data. AIC and BIC are based on different theoretical approaches to model comparison and have different assumptions and objectives. Consequently, there are different interpretations of the two criteria and also different possible results (Burnham & Anderson, 2004; Kuha, 2004). For the sake of interpretability, the BIC penalizes complex models (i.e., those with many parameters) to a greater extent than does the AIC (Wagenmakers & Farrell, 2004). However, as pointed by Kuha (2004), using the two criteria together is always advocated as agreement provides reassurance on the robustness of the results and disagreement still provides useful information for the discussion.

The aim of this paper is to introduce the model comparison approach considering an applied example. In this way, we can compare and discuss the different results obtained using the traditional NHST approach and the model comparison approach. The remainder of this paper is structured as follows. First, we introduce the case study regarding the evaluation of stereotype threat effects on Italian girls' mathematics performance. Subsequently, we present the analysis results considering separately the traditional NHST approach and the model comparison approach.

5.2 The Stereotype Threat Effects

As a case study, we evaluate the stereotype threat effects on Italian girls' mathematics performance. Many studies have found that males, on average, perform better than females in mathematics, although the size of this gender gap is small and varies considerably across countries (Fryer & Levitt, 2010). This difference may be due to the stereotype threat effect. Stereotype threat theory postulates a situational decrement in a person's performance owing to the awareness that his or her own ingroup is considered to be less skilful in the domain in which he or she is going to be tested (Spencer et al., 2016; Steele et al., 2002). Considering boys being better than girls in mathematics is a diffuse cultural stereotype since the elementary school (Nosek et al., 2009). Thus, stereotype threat has been proposed as an explanation for this gender gap in mathematics tests (Spencer et al., 1999).

In the literature, however, studies indicate contrasting results about the role of the gender stereotype threat effect. Thus, further analyses are required to evaluate whether and how a negative stereotype about women in mathematics impairs their performance.

5.2.1 The Study

To assess the presence or absence of the gender stereotype threat effect on boys' and girls' performance on mathematical problems, we evaluate participants' responses to mathematical problems before and after an experimental manipulation. This experimental manipulation aimed to elicit the gender stereotype that males are better than females in mathematics. Only half of the participants were assigned to the stereotype threat manipulation, whereas the other half were assigned to a control condition.

The study sample included 328 Italian students participants (155 females and 173 males) equally distributed according to school grade (ninth graders and eleventh graders). The first group (ninth graders) included 164 students (75 females and 89 males) with a mean age of 14.2 years. The second group (eleventh graders) included 164 students (80 females and 84 males) with a mean age of 16.2 years (see Table 5.1).

In the pre-test, participants responded to 18 mathematical problems. After the experimental manipulation, participants responded to another 18 different problems in the post-test. Mathematical problems for ninth graders and eleventh graders were different, thus the total number of unique problems was 72.

Table 5.1: Participants age according to gender, grade and condition ($n_{subjects} = 328$).

		Males		Females	
Condition	**Grade**	n	**Age Mean (SD)**	n	**Age Mean (SD)**
9th	NoST	44	14.2 (0.4)	36	14.2 (0.3)
	ST	45	14.1 (0.4)	39	14.2 (0.4)
11th	NoST	43	16.2 (0.5)	38	16.2 (0.5)
	ST	41	16.1 (0.3)	42	16.2 (0.4)

5.3 Statistical Analyses

First, we present the descriptive statistics. Subsequently, we conduct separate analyses following the traditional NHST approach and the model comparison approach, respectively. Note that in all statistical analyses the missing responses are considered to be wrong answers. All statistical analyses are conducted using the R statistical software (v4.1.0; R Core Team, 2021)

5.3.1 Descriptive Statistics

To compute mean accuracy and standard deviation, we first calculate the accuracy of each participant in the pre-test and post-test. Then we compute mean accuracy and standard deviations according to gender, grade and experimental condition (i.e., `ST` = stereotype threat condition; `NoST` = no stereotype threat condition). The values are reported in Table 5.2.

In Figure 5.1 the boxplots and violin plots of mean accuracy in the pre-test and post-test are presented as a function of gender, grade and experimental condition.

5.3.2 NHST Approach

To evaluate the presence of gender stereotype threat effect, we analyze participants' accuracy in the mathematical test in the different experimental conditions. Thus,

Table 5.2: Mean accuracy and standard deviations in the pre-test and post-test according to gender, grade and condition ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

		Males					Females				
			Pre-test		Post-test			Pre-test		Post-test	
Condition	**Grade**	***n***	**Mean**	**SD**	**Mean**	**SD**	***n***	**Mean**	**SD**	**Mean**	**SD**
NoST	9th	44	0.72	0.18	0.62	0.14	36	0.71	0.15	0.58	0.14
	11th	43	0.62	0.21	0.67	0.16	38	0.49	0.18	0.61	0.15
ST	9th	45	0.73	0.19	0.59	0.12	39	0.61	0.17	0.54	0.14
	11th	41	0.59	0.20	0.67	0.17	42	0.56	0.18	0.65	0.15

the dependent variable is the participants' responses to mathematical problems (`y`, `0` = wrong answer or missing response; `1` = correct answer).

Following the traditional NHST approach, we consider the *full model* that includes all the conditions present in our experimental design. Thus, the full model takes into account the four-way-interaction between:

- `condition`: `ST` = stereotype threat condition; `NoST` = no stereotype threat condition)
- `time`: `pre` = pre-test; `post` = post-test)
- `gender`: `M` = males; `F` = females
- `grade`: `9th` = ninth graders; `11th` for eleventh graders

We also include the random effects of the different participants (`ID`, $n_{subjects} = 328$) and problems used in the pre-test and post-test (`item`, $n_{items} = 72$) to take into account the individual variability of participants and mathematical problems. Thus, the resulting full model is a mixed-effects model and, using the R formula syntax, we have

```
y ~  condition * time * gender * grade + (1|ID) + (1|item)
```

Moreover, to take into account the characteristics of the dependent variable (binary outcome), we consider a generalized linear model (GLM), in particular a logistic regression model. This allows us to properly model the probability of correctly answering a mathematical problem.

Note that at this point we have already taken several decisions to properly model the data, instead of relying on a mindless application of statistical techniques. In

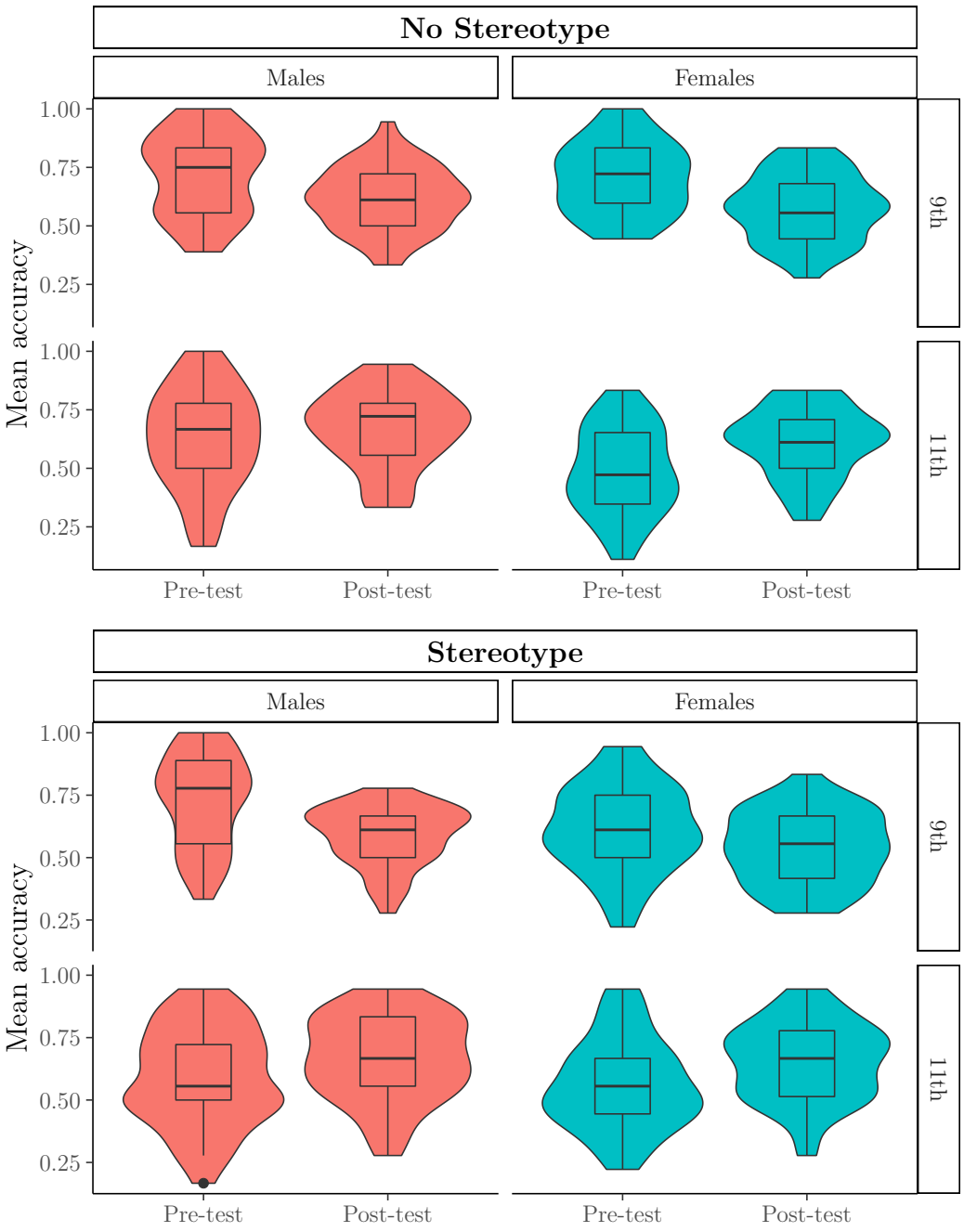

Figure 5.1: Boxplots and violin plots of mean accuracy in the pre-test and post-test as a function of gender, grade and experimental condition ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

this case linear models, as those assumed in both ANOVA and ANCOVA, are inappropriate statistical methods for analyzing stereotype threat effects. One fundamental problem is that the data are categorical (correct, incorrect, or missing responses to mathematics problems), and ANOVA and ANCOVA methods are not appropriate for categorical data analysis (Agresti, 2002) despite their widespread use in psychological research. Moreover, in the case of ANCOVA, a fundamental assumption is that the covariate is independent of the experimental effect, but group differences, such as a difference in the mathematical ability of boys and girls, violate that assumption. The variance explained by gender and the variance explained by mathematics ability cannot be separated, and spurious effects can arise. As Wicherts (2005) observed, "stereotype threat theory explicitly predicts violations of practically all assumptions underlying ANCOVA". These considerations should warn us against the mindless application of statistical techniques as black boxes, because they may lead to unreliable results.

After fitting the model using the R package `lme4` (Bates et al., 2014), we can run an analysis of *Deviance* using the `Anova()` function from the R-package `car` (Fox & Weisberg, 2019) to evaluate which are the important predictors in our full model. Results of the analysis of deviance are reported in Table 5.3. Note that, in the case of generalized linear models (GLM), the deviance is the corresponding of the residual variance used in the traditional ANOVA in the case of linear models.

Results indicate that the four-way interaction is statistically significant ($\chi^2(1) = 4.93$, $p\text{-}value = 0.026$). So, at this point, we would be rather happy because we have found a statistically significant result. Our experimental manipulation did actually work, but does this mean that we have found evidence for the stereotype effect? Actually to properly evaluate the stereotype effect we have to dig a little bit deeper into the model.

Evaluating the Stereotype Threat Effect

To evaluate the stereotype threat effect, we need to consider whether females in the stereotype condition performed worse in the post-test compared to females in the no stereotype condition. More precisely, we compare the difference between post-test and pre-test accuracy in the stereotype condition (ST) with the difference between post-test and pre-test accuracy in the no stereotype condition ($NoST$). Even more precisely, we are actually considering not the accuracy but its logit transformation. In fact, in logistic regression, accuracy (i.e., probability of correct response) is modelled after the logistic transformation is applied. Therefore, we have that:

$$ST_{effect} = (logit\_A_{ST;post} - logit\_A_{ST;pre}) - (logit\_A_{NoST;post} - logit\_A_{NoST;pre}),$$

where ST_{effect} is the stereotype threat effect, $logit\_A$ indicates the logit transformed accuracy, ST and $NoST$ indicate respectively the "Stereotype" and the "No Stereo-

Table 5.3: Analysis of Deviance of the full model ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

Effect	Chisq	Df	Pr(>Chisq)
Main Effects			
Condition	0.24	1	0.6246
Time	0.00	1	0.9574
Gender	12.40	1	0.0004
Grade	0.07	1	0.7901
Two-way Interactions			
Condition x Time	0.11	1	0.7417
Condition x Gender	0.21	1	0.6462
Time x Gender	3.15	1	0.0757
Condition x Grade	2.90	1	0.0885
Time x Grade	4.03	1	0.0446
Gender x Grade	0.06	1	0.8054
Three-way Interactions			
Condition x Time x Gender	0.25	1	0.6158
Condition x Time x Grade	0.00	1	0.9791
Condition x Gender x Grade	3.88	1	0.0488
Time x Gender x Grade	0.35	1	0.5564
Four-way Interaction			
Condition x Time x Gender x Grade	4.93	1	0.0265

Note:
Result from function `car::Anova()` with option `type = "II"` to respect the *"principle of Marginality"* (see help page `?car::Anova()` or Fox, 2016)

type" experimental conditions, and *post* and *pre* indicate the pre-test and post-test conditions. Note that the stereotype threat effect has to be evaluated considering the participants' performance in the different conditions within the same gender group. It would make no sense to consider the performance differences between gender groups (i.e., females and males). It could be possible that the stereotype threat also affects males, for example leading to an improvement in performance, however, this is not the focus of the analysis. Therefore, we consider only the comparisons within females, ignoring males.

To properly evaluate these planned comparisons, we need to define the appropriate contrasts. First, we generate the required model matrix with all the experi-

mental conditions of interest. As an example, consider the following code:

```
# Define possible experimental conditions
my_cond <- expand.grid(gender = c("M","F"),
            grade = c("9th","11th"),
            condition = c("NoST","ST"),
            time = c("pre", "post"))

# Obtain Model Matrix
mm <- model.matrix(~ condition * time * gender * grade, data= my_cond)

# Create labels of type "F_9th_ST_post"
rownames(mm) <- apply(my_cond, MARGIN = 1, paste, collapse = "_")

# Define contrasts of interest
my_contrast <- rbind(
  "diff_9th" = (mm["F_9th_ST_post",] - mm["F_9th_ST_pre",]) -
    (mm["F_9th_NoST_post",] - mm["F_9th_NoST_pre",]),
  "diff_11th" = (mm["F_11th_ST_post",] - mm["F_11th_ST_pre",]) -
    (mm["F_11th_NoST_post",] - mm["F_11th_NoST_pre",])
)
```

Subsequently, we define and test the contrasts using the R-package `multcomp` (Hothorn et al., 2008). Results are reported in Table 5.4.

Table 5.4: Planned comparisons to evaluate stereotype threat effect in 9th grade and 11th grade females ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

	Hypotheses	**Estimates**	**S.E.**	**Z value**	**p-value**
diff_9th	diff_9th == 0	0.273	0.179	1.529	0.252
diff_11th	diff_11th == 0	-0.124	0.179	-0.696	0.973

Note: Adjusted p-values reported using Bonferroni correction.

We can observe that there is no statistically significant difference for either 9th-grade or 11th-grade females. Broadly speaking, these results suggest that there is no statistical evidence of the presence of the stereotype threat effect (note that, strictly speaking, we can actually only state the there was no evidence against the null hypothesis). Note that in both cases, a two-tailed test was conducted and the p-value was adjusted according to Bonferroni correction to account for multiple comparisons.

Given that there is no stereotype threat effect, how should we interpret the four-way interaction? Probably it is not immediately clear, but a model with a four-way interaction is a kind of monster model with a very large number of parameters. All the parameters of the full model are reported in Table 5.5.

Table 5.5: Estimated parameters of the full model ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

Parameter			95% CI			
Name	**Estimate**	**S.E.**	**lower**	**upper**	**Z value**	**p-value**
Main Effects						
Intercept	1.13	0.30	0.54	1.71	3.80	0.0001
Condition (Stereotype)	0.02	0.20	-0.36	0.41	0.13	0.8998
Time (Post)	-0.50	0.39	-1.26	0.26	-1.29	0.1984
Gender (Female)	-0.09	0.21	-0.49	0.32	-0.41	0.6797
Grade (11th)	-0.45	0.42	-1.27	0.37	-1.08	0.2796
Two-way Interactions						
Condition x Time	-0.19	0.17	-0.52	0.14	-1.16	0.2476
Condition x Gender	-0.51	0.29	-1.08	0.05	-1.77	0.0764
Time x Gender	-0.14	0.18	-0.49	0.21	-0.78	0.4381
Condition x Grade	-0.22	0.28	-0.77	0.33	-0.77	0.4398
Time x Grade	0.82	0.55	-0.26	1.90	1.49	0.1355
Gender x Grade	-0.66	0.29	-1.23	-0.09	-2.27	0.0231
Three-way Interactions						
Condition x Time x Gender	0.47	0.25	-0.01	0.95	1.91	0.0566
Condition x Time x Grade	0.38	0.25	-0.10	0.87	1.57	0.1174
Condition x Gender x Grade	1.10	0.41	0.31	1.90	2.71	0.0067
Time x Gender x Grade	0.51	0.25	0.01	1.00	2.00	0.0455
Four-way Interaction						
Condition x Time x Gender x Grade	-0.78	0.35	-1.47	-0.09	-2.22	0.0265
Random Effects						
ID (Intercept)	0.73					
Item (Intercept)	1.11					

Note: Baseline category for condition is "No Stereotype". Baseline category for time is "Pre". Baseline category for gender is "Male". Baseline category for grade is "9th". Confidence intervals computed using the Wald method (not availble for the random effects as the likelihood function is not symmetrical).

To interpret the four-way interaction, the predicted mean accuracy and confidence intervals in the pre-test and post-test as a function of gender, grade and experimental condition are presented in Figure 5.2.

From a descriptive point of view, we can observe that 9th-grade males and females had lower accuracy in the post-test than in the pre-test in both experimental conditions (i.e., "No Stereotype" and "Stereotype"). On the contrary, 11th-grade

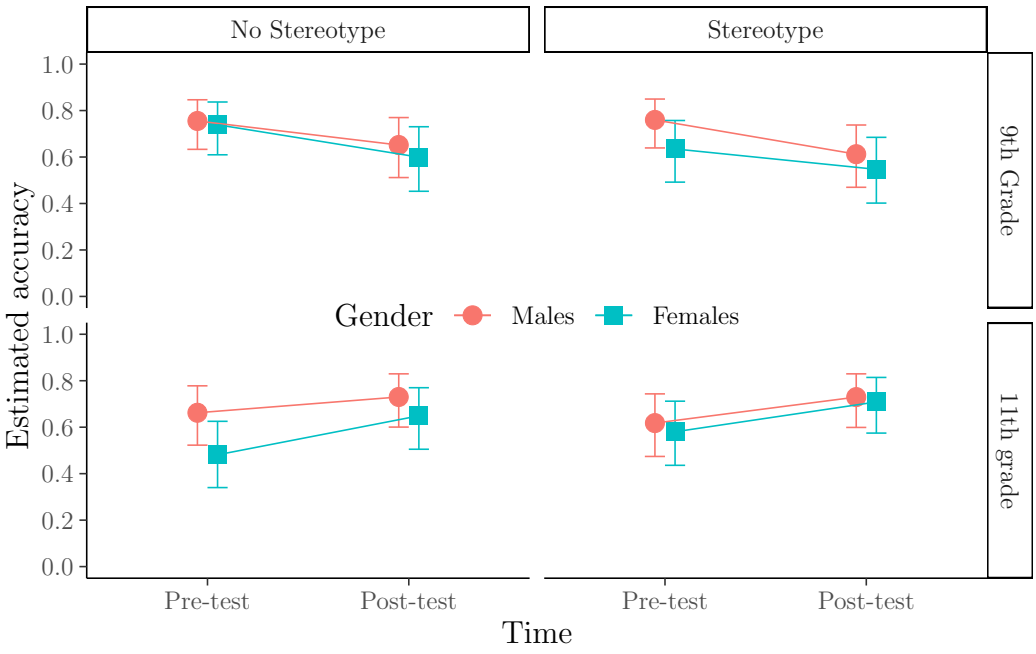

Figure 5.2: Predicted accuracy in the pre-test and post-test as a function of gender, grade and experimental condition ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

males and females had higher accuracy in the post-test than in the pre-test in both experimental conditions (i.e., "No Stereotype" and "Stereotype"). These patterns are likely due to differences in the difficulty of the four tests rather than the experimental manipulation.

Surely, there are some statistically significant differences between the experimental conditions, but to find them we should test all the possible comparisons. This, however, leads to a broader consideration: are we evaluating some important and theoretically supported phenomena or are we just modelling the random noise? Without some clear hypotheses, testing all possible comparisons may lead to unreliable results. Therefore, we do not proceed any further in the analysis but we rather move to the model comparison approach.

5.3.3 Model Comparison Approach

As previously introduced in Section 5.1.1, in the model comparison approach, first we need to define the different models according to our research hypothesis. Subsequently, we can compare the obtained models using the information criteria.

Model Definitions

To evaluate the presence of gender stereotype threat effect, we compare 7 different mixed-effects logistic regression models for the same reasons as described in Section 5.3.2. In all models, the dependent variable `y` is accuracy (`0` = wrong answer or missing response; `1` = correct answer), and `ID` and `item` are included as random effects to account for the variability of participants and mathematical problems.

In model `m0` we only consider the random effect of participants ($n_{subjects} = 328$) and the random effect of items ($n_{items} = 72$). This model is used as a reference model to evaluate the possible contribution of the other variables.

Next, we evaluate the possible contribution of `gender` and `grade`, defining different combinations of the two variables in models `m1`, `m2`, and `m3`. In model `m1` we add only the effect of `gender` to model `m0`. This model evaluates whether gender is the only important variable that explains accuracy in both the pre-test and post-test. Model `m2` includes the additive effect of `gender` and `grade`. This model evaluates whether there is also a role of grade in addition to the gender effect. Any effect of grade on accuracy could be related to differences in children's age, differences in the difficulty of the test problems used for the two grades, or both. Model `m3` adds the interaction between `gender` and `grade` to evaluate whether the effect of grade plays a different influence on mathematical accuracy depending on gender.

Next, we evaluate possible effects of the experimental manipulation. Model `m4` adds the interaction between `time` and `condition` to the effects of `gender` and `grade`. This model evaluates whether the experimental manipulation had an overall effect that remained constant independently of gender and grade. Model `m5` includes the three-way interaction between `time`, `condition` and `gender`. This model is consistent with the hypothesis that the stereotype threat effect (i.e., the experimental manipulation) should differently affect boys and girls. Finally, Model `m6` includes the four-way interaction between test `time`, `condition`, `gender`, and `grade` to evaluate whether the experimental manipulation could have differently influenced subjects according to gender and grade.

The models are summarized below using the R formula syntax:

```
m0 : y ~ 1 + (1|ID) + (1|item)
m1 : y ~ gender + (1|ID) + (1|item)
m2 : y ~ gender + grade + (1|ID) + (1|item)
m3 : y ~ gender * grade + (1|ID) + (1|item)
m4 : y ~ condition * time + gender + grade + (1|ID) + (1|item)
m5 : y ~ condition * time * gender + grade + (1|ID) + (1|item)
m6 : y ~ condition * time * gender * grade + (1|ID) + (1|item)
```

Considering the research hypotheses, `m1` and `m5` are of particular interest as they support, respectively, the absence or the presence of the stereotype threat effect. If the stereotype threat effect is present, and differently affected boys and girls, we expect `m5` to be selected as the best model; whereas if only the gender difference in mathematical abilities is supported by the data, then we expect `m1` to be the best model. The other models allow us to consider the possible effects of other variables and their interactions.

Model Comparison Results

After estimating the models, the AIC and BIC values together with their relative weights are computed[2]. Results are reported in Table 5.6.

Table 5.6: Model comparison using AIC and BIC ($n_{subjects} = 328$; $n_{Items} = 72$; $n_{observations} = 11808$).

Model	Df	AIC	$AIC_{weights}$		BIC	$BIC_{weights}$	
m0	3	13099.85	0.00		13121.98	0.21	
m1	4	13089.84	0.63		13119.34	0.78	
m2	5	13091.78	0.24		13128.66	0.01	
m3	6	13093.71	0.09		13137.97	0.00	
m4	8	13097.42	0.01		13156.44	0.00	
m5	11	13099.78	0.00		13180.92	0.00	
m6	18	13097.89	0.01		13230.67	0.00	

Model `m1` is the most likely model given the data and the set of models considered. It is interesting, however, to observe the second-best model according to the AIC and BIC. According to the AIC, `m1` is the best model with 63% probability and `m2` is second best with 24% probability. According to the BIC, `m1` is the best model with 78% probability and `m0` is second best with 21% probability. In both cases the probability of `m5` is infinitesimal.

[2]Relative weights represent the relative likelihood of each model. This can be interpreted as how likely each model is to be the best model among the set considered models

AIC tends to select more complex models that can better explain the data, and in this case, it does not completely exclude model `m2` that includes the role of grade. On the contrary, BIC penalizes complex models to a greater extent than AIC, and in this case, the probability of `m2` is only 1% and the probability of the null model `m0` increases to 21%. Overall, given that both AIC and BIC select `m1` and give infinitesimal probability to `m5`, the results support the hypothesis that there is only a gender effect and no evidence for a stereotype threat effect. We can now take a deeper look into the selected model.

Description of Model `m1`

Model `m1` includes only the effect of `gender` and the random effects of `ID` and `item`.

$$\texttt{m1} : \texttt{y} \sim \texttt{gender + (1|ID) + (1|item)}$$

The estimated parameters of the model `m1` are reported in Table 5.7.

Table 5.7: Estimated parameters of the selected model ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

Parameter			95% CI			
Name	**Estimate**	**S.E.**	**lower**	**upper**	**Z value**	**p-value**
Main Effects						
Intercept	0.82	0.15	0.52	1.11	5.46	0.0000
Gender (Female)	-0.33	0.09	-0.51	-0.14	-3.50	0.0005
Random Effects						
ID (Intercept)	0.74		0.67	0.83		
Item (Intercept)	1.14		0.97	1.37		

Note: Baseline category for gender is "Male". Confidence intervals computed using the profile likelihood.

Predicted mean accuracy and confidence intervals according to gender are presented in Figure 5.3 and Odds Ratios (OR) for accuracy are reported in Table 5.8. Overall, results indicate that males perform better than females on mathematical problems. However, the difference is small (males predicted accuracy = .70; females predicted accuracy = .62).

Model fit

Note that information criteria do not provide an absolute measure of the quality of a model, but only a relative measure used to compare the different models. Thus,

Figure 5.3: Predicted accuracy according to gender ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

Table 5.8: Odds Ratios of the selcted model ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

		95% CI	
Parameter	**OR**	**lower**	**upper**
Intercept	2.26	1.68	3.04
Gender (Female)	0.72	0.60	0.87

Note: Baseline category for gender is "Male".

the model comparison indicates to us which are the most likely models relative to the set of models considered, but it does not allow us to evaluate whether these models properly fits the data.

To evaluate the fit of model `m1` to the data, we consider the R^2, which, unlike the information criteria, evaluates the absolute value of the goodness-of-fit of a model. In the case of generalized mixed-effects models, however, there are several definitions of R^2. We compute the *Marginal* R^2 and the *Conditional* R^2 as suggested by Nakagawa and Schielzeth (2013). They explain that *Marginal* R^2 is concerned with the variance explained by fixed factors of the model, and *Conditional* R^2 is concerned with the variance explained by both fixed and random factors of the model. *Marginal* R^2 and the *Conditional* R^2 are computed using the R package

`MuMIn` (Bartoń, 2019). Values are reported in Table 5.9.

Table 5.9: Selected model Marginal R^2 and Conditional R^2 ($n_{subjects} = 328$; $n_{items} = 72$; $n_{observations} = 11808$).

Marginal R^2	**Conditional R^2**
0.005	0.364

Overall, the model presents a good fit to the data. However, this result is mainly given by the inclusion of the random effects, whereas the fixed effect related to gender differences contributes in a really limited way.

5.4 Conclusions

In this paper, we compared two different statistical approaches: the Null Hypothesis Significance Testing (NHST) approach and the model comparison approach using information criteria.

The NHST has developed into a mechanical application of statistical testing by always considering the catch-all null hypothesis that "*nothing is going on*", rather than properly formalize and evaluate the hypotheses of interest. This mindless application of the NHST can easily lead to unreliable results as statistically significant results may occur even just by chance for many different factors other than the presence of a real effect. The researcher degrees of freedom, large measurement errors, and small sample sizes all contribute to the creation of noise in the data and the subsequent misleading results. Moreover, when conducting a study, researchers usually want to quantify the evidence in favour of their hypotheses. Despite its popularity, however, the NHST does not allow us to do that, but it only allows us to quantify the evidence against the null hypothesis. Unfortunately, this subtle but important difference is often neglected by researchers favouring an erroneous interpretation of the results.

On the contrary, the model comparison approach allows to directly compare different hypotheses and evaluate the relative evidence in favour of one hypothesis according to the data. Moreover, the model comparison approach favours the formalization of research hypotheses into appropriate statistical models that reflect the data generative process and allows the selection of models with effects that offer an appropriate description of the data, penalizing the inclusion of further, unnecessary effects that only introduce an unnecessary level of complexity.

Considering the Stereotype Threat effect example, the results are very indicative. Following the NHST approach, we would have ended up interpreting some strange,

unexpected differences within a four-way interaction without realizing that, probably, we were just modelling the random noise in the data. Following the model comparison approach, instead, we could directly evaluate our hypotheses of interest obtaining a straightforward interpretation of results: there is no evidence for the stereotype threat effect but only for gender differences.

Of course, the differences in the results obtained with the two approaches in the stereotype threat effect example are very remarkable. In other cases, however, we could easily expect the two approaches to lead to similar conclusions. This would not be a surprise and, actually, we are not trying to suggest that one approach is better than the other. Both are just statistical approaches that when properly applied to answer the right question will provide reliable results. Issues arise when a mindless application of statistical techniques occurs. Hopefully, at this point, it should be clear that statistical inference is a complex process that involves several decisions (i.e., the famous researcher degrees of freedom). It is important, therefore, to be aware of all pros and cons of the different statistical methods to consciously use them in different situations, avoiding the application of statistical techniques as black boxes. In this regard, however, we think that the model comparison approach enhances researchers' reasoning about statistical inference more than the mechanical application of the NHST approach. For this reason, we hope that in the future there will be less speaking about testing and more about modeling.

Round Table

1. Chapter 5 makes the general point that model comparison can somehow solve some of the issues of NHST. However, one can also argue that the model comparison approach offers a larger number of researchers' degrees of freedom than an NHST approach.

   Answer: We use this point as an opportunity to further discuss the differences between model comparison and NHST. To do that, let's consider two different aspects of statistical inference: model definition and hypothesis testing (this is not a complete discussion on the topic, for more details see Fox, 2016; McElreath, 2020b):

   (a) **Model Definition:** Everything starts from the definition of a statistical model. The model definition could be fully explicit (as in the model comparison approach) or less explicit (as in the NHST), however, we always need to define a statistical model. Thus, only apparently there are a larger number of researchers' degrees of freedom in the model comparison approach than in the NHST approach. Actually, we have the same degrees of freedom because the definition of a statistical model is always required. Even in the case of ad-hoc test procedures (e.g., t-tests, ANOVA, ANCOVA, MANOVA), we are still assuming a specific statistical model, see blog post: https://lindeloev.github.io/tests-as-linear/ "*Common statistical tests are lin-*

ear models (or: how to teach stats)". Thus, one of the advantages of model comparison is that the models' definition is explicit. This forces researchers to reason about the model definition thinking about the data generative process (is the selected distribution adequate? Do you expect relations other than linear?). In doing that, the researchers have many degrees of freedom (degrees of freedom that the researchers would have in any case), but now researchers have to properly justify their choices as everything becomes explicit instead of remaining implicit in some mechanical application of statistical techniques. Of course, it is possible to argue about specific choices but this is a positive aspect as discussion about relevant issues could lead to model improvements.

(b) **Hypothesis Testing:** In the NHST approach, hypotheses are defined as constraints on a parameter of interest (or a subset of parameters usually set at 0) of the full model. Next, given an appropriate test statistic (e.g., t-test, F-test, or Likelihood-ratio test), we consider the probability of obtaining more extreme results under the Null-hypothesis than the observed ones (i.e., p-value). In the model comparison approach, instead, different models are formalized according to different hypotheses. Next, using information criteria, these models are compared according to their ability to predict new data (i.e., out-of-sample deviance).

Strictly speaking, in model comparison with information criteria there is no statistical test (no p-value is computed) but rather we are assessing a characteristic of the model as a whole. Thus, rather than hypothesis testing, it would be more correct to discuss model comparison within the more general framework of model selection. Model selection is a process that selects a statistical model from a set of candidate models, given the data. This sounds the same as model comparison and, in fact, it is. More precisely model comparison is a specific type of model selection. The difference between model comparison and model selection is how the set of candidate models is defined. In model comparison, the idea is that the different models formalize relevant theoretical perspectives that the researchers are interested to compare, thus models should be defined according to relevant scientific hypotheses and principles and they should have a meaningful interpretation. In model selection, instead, models can be defined in any way, for example, an automatic procedure of variable selection as in the case of stepwise analysis. In this case, however, we could end up with models that have no clear interpretation. Thus, in the model comparison approach, we have a limited number of relevant models, whereas in model selection we have a larger set of models.

Note that model selection (as well as model comparison) can be done according to different criteria. We can even use statistical hypothesis tests for model selection. The NHST approach, however, has several limitations: we can only reject the null hypothesis and not evaluate the evidence in favour of a model; given a sufficiently large sample, even trivial effects will be statistically significant; we can only test nested models. On the contrary, information criteria allow us to assess evidence in favour of a model, penalize for model complexity (remove trivial effects), and compare non-nested models. Thus, as long as models are defined according to hypotheses of interest, we can use information criteria to *test* our hypotheses. One of the limits of information criteria, however, is that we can not define inequality

constraints on the model parameters. In the NHST approach, we can conduct directional tests but no analogous exists with the information criteria. To overcome this issue, we need to move on to the Bayes factor, presented in Chapter 6.

2. I have the feeling that sometimes the difference between Model Comparison and NHST (and testing in general) is overemphasized. In fact, to some extent, NHST can be seen as a form of simplified model comparison, for instance, when you compare a model to its null version (e.g., you test for the interaction by comparing a model lm(y a + b + a*b) with lm(y a + b)). In light of this consideration, the Candidate should make very clear since the beginning (pages 5-6) that the model comparison approach, taken as a whole, can provide a more nuanced picture and provides a more versatile tool for testing hypotheses that may not be tested under an NHST approach. Otherwise, the distinction between the two approaches cannot be appreciated enough.

   Answer: In the chapter, we introduced the model comparison as an alternative approach to the NHST emphasizing their differences. As pointed out by the reviewer, however, this is a simplified (and to some extend incorrect) simplification. We discussed in the previous point how model comparison is just as a particular type of model selection. To select the preferred model among a set of candidate models, we need to compare them according to some criteria. One of the most common criteria is actually the statistical hypothesis tests, considering for example the likelihood ratio test. The NHST approach, however, has several limitations: we can only reject the null hypothesis and not evaluate the evidence in favour of a model; given a sufficiently large sample, even trivial effects will be statistically significant; we can only test nested models. On the contrary, information criteria (or the Bayes factor) allow us to assess evidence in favour of a model, to penalize for model complexity (remove trivial effects), and to compare non-nested models. Thus, rather than considering NHST and model comparison as two different approaches, it is more correct to consider statistical significance, information criteria, and the Bayes factor as different criteria for model selection.

3. "For this reason, we hope that in the future there will be less speaking about testing and more about modeling" (p.78). OK, but the two terms are not mutually exclusive. Bayesian hypothesis testing approaches such as Bayes Factor rely on the model comparison, as the Candidate nicely shows in the very next Chapter.

   Answer: We agree with the reviewer that modeling and testing are not mutually exclusive. Actually, as we discussed in a previous point, these are two different stages of statistical inference. Usually, there is a narrow focus on testing overlooking the modeling part. However, the definition of appropriate statistical models is required to obtain reliable results when testing research hypotheses. Thus, we hope for increasing awareness about the importance of appropriate modeling before testing.

4. The dissertation uses very often the word "mindless" to refer to the mechanical application of NHST. But it is quite obvious that doing anything "mindlessly" in research does not lead to anything good, and referring that mostly to the NHST might contribute to creating a strawman. For example, in Section 5.3.2, "Note

that at this point ...". The paragraph seems to suggest that someone out there is (mindlessly) analyzing dichotomous outcomes, without even computing average accuracy scores, with simple linear models. Fortunately, I do not think this is the case (at least, I never met anybody doing that!). Perhaps, the use of "mindless" might be reduced, or clearer examples of such practices (e.g., in the stereotype threat literature) should be mentioned.

Answer: We agree with the reviewer that repeatedly referring to the mechanical application of the NHST as a "*mindless*" procedure could create a straw-man. However, we wanted to emphasize that the mechanical application of any statistical procedures is never appropriate. Researchers should always question whether a given statistical approach is appropriate and what are its limits. For example, analysing dichotomous outcomes computing average accuracy scores we lose information about the total number of trials (50% of accuracy could be obtained as 1 out of 2 or 5 out of 10); using simple linear models instead of logistic regression cold lead to predictions out of the 0-1 range. Of course, these are stupid examples that never happen in reality (hopefully). However, we want to highlight the importance of carefully thinking about statistical analyses. Statistical reasoning is not a mechanical application of statistical procedures and the best thing to do may not be what we were used to doing.

5. "Monster model" (p.71) does not sound very scientific to me. How would you define a "monster model"?

   Answer: It is true "*monster model*" does not sound very scientific [it is, actually, an implicit reference to the famous Golem of Prague described by McElreath (2020b). In psychology, it is common to consider 3-way, 4-way, or even higher-order interactions. By doing this, however, researchers may not be aware that the number of parameters in the model increases extremely rapidly and they could end up creating a "*monster model*".

6. "Results indicate that the four-way interaction is statistically significant ($\chi^2(1) = 4.93, p\text{-}value = .026$). So, at this point, we would be rather happy because we have found a statistically significant result. Our experimental manipulation did actually work, but does this mean that we have found evidence for the stereotype effect? Actually to properly evaluate the stereotype effect we have to dig a little bit deeper into the model" (p.68). I think this is a misrepresentation of what a decently trained scholar would infer from these results. It is a four-way interaction, and it should be interpreted as a different three-way interaction conditional on the levels of one of the independent variables. This does not automatically lead to concluding that the stereotype threat works. In fact, this finding only leads to the observation that stereotype threat interacts with gender groups and time differently across grade groups.

   Answer: In the chapter, we oversimplified the interpretation of the results to emphasize the feeling of "*I found a statistically significant result*". But as pointed out by the reviewer, this is a misrepresentation of the results. We thank the reviewer for clarifying this point.

7. "Therefore, we consider only the comparisons within females, ignoring males" (p.69). But is this justified, since the experiment was set up to test a 4-way interaction? Shouldn't one consider a (quite complex) contrast including also males?

   Answer: This is indeed a simplification of the problem but the rationale behind this is the following. The stereotype threat effect assumes a decrement in a person's performance due to the activation of the stereotype that his or her own ingroup is considered to be less skilful. In our case, girls' performance in mathematics is expected to decrease due to the activation of the stereotype "*boys are better than girls in mathematics*". Thus, to properly evaluate females' decrement in performance, we have to compare the difference between pre-post females' performance in the stereotype condition and in the control condition. This gives us a measure of the stereotype effect in females. As the reviewer pointed out, in this case we are ignoring males. However, this is justified as we do not need to consider males' performance to assess the stereotype effect in females. In both cases, males' and females' performance differences are computed within gender, thus, males' performance is not required to evaluate females' decrement. Considering males' performance, we could find no differences, an increment in performance or a decrement (as expected in the females' group), but this would not affect the evaluation of females' decrement. Evaluating together females' and males' performance is important to properly interpret the results from a theoretical perspective (in case of a decrement in both groups how would you interpret the stereotype threat effect?). To evaluate the stereotype threat effect in females', however, we only need to consider females' contrasts. Note that this was the main focus of the study and also what we would expect according to the theoretical definition of the stereotype threat effect. Of course, evaluating the differences with males is important to have a general comprehension of the phenomenon but it is not necessary for our aims.

8. I don't understand why the discussion focuses on model 5 rather than on model 6, which was the one that was preferred by the NHST approach.

   Answer: Model 5 is consistent with the hypothesis that the stereotype threat effect affects boys and girls differently. Model 6, instead, is consistent with the hypothesis that the stereotype threat effect affects subjects differently according to gender and grade (i.e., stereotype threat could be present only in older subjects). In the NHST approach, we focused on model 6 because it is the maximal model according to the experimental design. However, the main interest was to evaluate the presence of the stereotype threat effect (age differences were explored only as a possibility). Thus, in the discussion, we focused on model 5.

9. It would be useful to understand whether the problems presented pre and post-intervention can be considered parallel tests from a psychometric standpoint.

   Answer: Items were carefully selected to obtain equal test difficulties for the two grade groups. However, this could not be guaranteed a priori. Thus, the grade was included to control for possible differences in test difficulties and, in model 6, to evaluate whether stereotype threat effect could differ according to age.

10. "These patterns..." (p.72). This interpretation shows the fact that not ensuring that test difficulties were the same across grades and time points may potentially threaten the validity of your conclusions.

    Answer: We agree with the reviewer that, ideally, test difficulties should have been the same across grades and time points. Items were carefully selected for this aim but, unfortunately, we could observe some differences in the difficulty of the four tests. However, we think that validity of the results is still solid as there is no floor or ceiling effect in any of the four tests, and the inclusion of control groups in the study design (both for females and males) guarantee to control for differences in the difficulty of the four tests.

11. I think that, on top of random intercepts, random slopes for the within-subjects effects (namely times) should have been modeled to follow the maximal structure recommendation from Barr et al. (2013).

    Answer: As pointed out by the reviewer, random slopes could have been included in the model. However, we defined a more parsimonious random structure, considering only subjects and items variability, for the following reason. When including random effects, model comparison is very likely to choose as the preferred model one with the most complex random structure (in psychology, there is always some individual variability in the effects). Thus, if we are only interested in evaluating the role of fixed effects, it is desirable to keep the same random structure for all models. In our case, we should have included random slopes in all models, also in those without the fixed effect of `time`. This appeared not reasonable to us, therefore we decided to include only random intercepts for `ID` and `item` in all models. Moreover, considering the debate regarding maximal vs. parsimonious structure, we point out Bates et al. (2018) response to Barr et al. (2013) recommendation, in which issues related to overparameterization of random structure are discussed (e.g., failure to converge and uninterpretable models).

6

Evaluating Informative Hypotheses with Equality and Inequality Constraint: The Bayes Factor via Encompassing Prior Approach[1]

Road Map

In the previous chapters, we introduced the model comparison approach using the information criteria to evaluate research hypotheses. Information criteria, however, do not allow to compare informative hypothesis with inequality constraints. In this chapter, we introduce the model comparison approach using the Bayes Factor with encompassing prior approach that allows us to properly evaluate informative hypotheses with equality and inequality constraints. As a case study, we consider the evaluation of attachment theories regarding the role of mother and father attachment on children's social-emotional development.

[1]This chapter is an original work in collaboration with Marci, T., De Carli, P., and Altoè, G. I contributed to conceiving the original idea, writing of the manuscript, statistical analysis and the graphical representations. Supplemental Materials available at https://claudiozandonella.github.io/Attachment/. GitHub repository https://github.com/ClaudioZandonella/Attachment.

6.1 Introduction

When conducting a study, researchers usually have expectations based on hypotheses or theoretical perspectives they want to evaluate according to the observed data. In fact, the evaluation of research and theoretical hypotheses is one of the principal goals of empirical research.

In psychology, the dominant statistical approach to evaluate research hypotheses is the Null Hypothesis Significance Testing (NHST). In the literature, however, the utility and validity of NHST are largely debated (Wasserstein et al., 2019). This approach presents indeed several limitations. First, the NHST places a narrow focus on statistical testing rather than on the formalization of hypotheses. This has usually led researchers to evaluate data against the catch-all null-hypothesis that nothing is going on rather than testing their specific expectation (and the alternative hypothesis is rarely formalized). Second, the p-value does not quantify the evidence in favour of one hypothesis, thus, it is not possible to *"accept"* a hypothesis but only to *"reject"* it. This is inconvenient as, in case of not significant results, the researchers are left in a state of indecision. Third, NHST does not allow testing multiple hypotheses at the same time. Using the NHST the null hypothesis is tested only against a single alternative. Fourth, NHST is unsuitable for testing broad classes of hypotheses with equality and inequality constraints (Mulder & Olsson-Collentine, 2019). Expectations can be evaluated using one-side tests, however, when more groups or more variables are involved, it is not possible to evaluate complex parameter constraints that reflect researchers' expectations (van de Schoot et al., 2011).

Model comparison is a different approach that allows researchers to compare multiple hypotheses and identify which is the most supported by the data (McElreath, 2020b). Hypotheses are first formalized as statistical models according to researchers expectations or theoretical perspectives. Subsequently, it is possible to evaluate which is the most supported model among those considered according to the data. To do that a popular approach is to use information criteria such as the AIC or BIC criteria that estimate models ability to predict new data (Akaike, 1973; Schwarz, 1978; Wagenmakers & Farrell, 2004). Model comparison has several advantages and, in particular, it allows to directly compute the relative plausibility of each model given the data and the set of models considered. Model comparison using the information criteria, however, is not appropriate to evaluate hypotheses that include complex parameter constraints reflecting researchers' expectations. In fact, information criteria evaluate models complexity according to the number of parameters, but they do not take into account possible order constraints on the parameters.

An alternative criterium to evaluate research hypotheses in a model comparison is the Bayes Factor. In the last 25 years, there has been an increasing interest in the

Bayes Factor and its use has been proposed as the solution to the critical issues of the NHST (Heck et al., 2020; Mulder & Wagenmakers, 2016). Although it has its own limitations and it does not solve the fundamental issues of the misuse of statistical techniques (Gelman et al., 2013; Schad et al., 2021), the Bayes Factor offers some clear advantages. In particular, the Bayes Factor allows us to compare hypotheses obtaining a relative index of evidence, like the information criteria, but it also allows us to easily compare complex research hypotheses. In fact, so-called *"informative hypotheses"* (i.e., hypotheses containing information about the ordering of the model parameters) can be formalized according to researchers' expectations or theoretical perspectives using equality and inequality constraints (van de Schoot et al., 2011). Subsequently, these hypotheses can be compared against each other using the Bayes Factor.

The evaluation of informative hypotheses is of particular interest to the researchers as it allows them to directly assess specific complex expectations and theoretical perspectives. In the literature, a particular approach allowing to easily compute the Bayes Factor with complex informative hypotheses has received increasing attention: *"the Bayes Factor with encompassing prior"*. van de Schoot et al. (2011) presented a general introduction to informative hypotheses testing using the Bayes Factor with the encompassing prior approach, whereas Hoijtink (2012) offered a more detailed description of the development of this method. Other studies, instead, considered the application of this approach with specific statistical models; for example: mixed effect models (Kato & Hoijtink, 2006), evaluation of correlation coefficients (Mulder, 2016) and multinomial models (Heck & Davis-Stober, 2019). In addition, Gu et al. (2014) proposed a general approximate procedure to evaluate inequality constraints in a wide range of statistical models.

All studies mentioned above, however, considered only informative hypotheses with inequality constraints. Evaluating in the same hypothesis equality and inequality constraints, instead, it was no possible and equality constraints had to be approximate to "*about equality constraints*" (i.e., a equality constraints of type $\theta_i = \theta_j$ is approximated to $|\theta_i - \theta_j| < \xi$ for a value of ξ small enough).

Only recently, Gu et al. (2018) introduced an approximate procedure to evaluate both equality and inequality constraints using the Bayes Factor with the encompassing prior approach in a wide range of statistical models (i.e., generalized linear mixed models and structural equation models). Subsequently, this approach was extended by Mulder and Gelissen (2019) to generalized multivariate probit models and, finally, Mulder and Olsson-Collentine (2019) proposed an accurate procedure (i.e., not based on an approximation) that allows testing informative hypotheses with equality and inequality constraints in linear regression models.

The development and implementation of this approach are of particular interest because, although with its limits, it allows researchers to directly evaluate their

expectations and hypotheses. The available literature, however, is rather technical. The complexity of these articles makes it difficult for researchers not familiar with this approach to clearly understand all the steps involved. On the other hand, articles offering a general introduction to the Bayes Factor do not provide enough details to allow readers to autonomously apply this method to their problems, but they usually rely on ad-hoc solutions implemented in some statistical software.

The aim of this paper, therefore, is to offer a clear and detailed description of the Bayes Factor with the encompassing prior approach to allow other researchers to apply this approach in their studies. In particular, we refer to the approximated method proposed by Gu et al. (2018) as it applies to a wider range of conditions than the more accurate approach proposed by Mulder and Olsson-Collentine (2019). The paper is organized as follows. First, the method is introduced providing a detailed description of all steps and elements involved in the formalization of informative hypotheses and Bayes Factor computation. Subsequently, an application of the method to real data is presented to offer the opportunity to discuss the typical issues encountered in real complex scenarios.

6.2 Bayes Factor for Informative Hypothesis Testing

The evaluation of informative hypotheses with equality and/or inequality constraints involves several steps and elements. In this section, first, we describe how to formulate informative hypotheses. Subsequently, we introduce the Bayes Factor considering the encompassing prior approach based on the approximated method proposed by Gu et al. (2018).

6.2.1 Formulation of Informative Hypothesis

Informative hypotheses can be defined according to researchers' expectation, evidence from the literature or theoretical perspectives and they are formed by equality and/or inequality constraints on certain model parameters. These constraints are obtained as a linear combination of certain parameters and eventual constant values. For example, it is possible to state that two parameters are equal ($\theta_i = \theta_j$), one parameter is greater than another ($\theta_i > \theta_j$), or express other complex conditions as the difference between two parameter is less than a given value ($\theta_i - \theta_j < .5$; $2 \times \theta_i - \theta_j < .1 \times \sigma$).

Thus, an informative hypothesis H_i with equality and inequality constraints can be expressed in the form

$$H_i : \quad R_E\boldsymbol{\theta} = r_E \quad \& \quad R_I\boldsymbol{\theta} > r_I, \tag{6.1}$$

where R_E is a matrix expressing the equality constraints and r_E is a vector containing the constant values of the equality constraints. Whereas, R_I is a matrix expressing the inequality constraints and r_E is a vector containing the constant values of the inequality constraints. Finally, $\boldsymbol{\theta}$ is a vector with the model parameters involved in the constraints.

As an example, consider a study evaluating the efficacy of a new psychological treatment that is supposed to improve a given cognitive ability. In the study, a control group receiving no treatment and another group receiving the traditional treatment were included as a comparison. Moreover, imagine that the new psychological treatment was administered in three different modalities to different groups: individually, in-group, and online. Researchers expect no differences between the three modalities but they hypothesize that the new treatment will perform better than the traditional one and this, in turn, will be better than the no-treatment control condition. We can express this hypothesis as

$$H_i: \quad \theta_{control} < \theta_{traditional} < \theta_{individual} = \theta_{group} = \theta_{online},$$

were $\theta_{control}$ is the parameter estimating the average score of the no-treatment control group, $\theta_{traditional}$ is the parameter estimating the average score of the group receiving the traditional treatment, and $\theta_{individual}$, θ_{group}, and θ_{online} are the parameters estimating the average scores of the groups receiving the new treatment individually, in group, and on-line, respectively. The corresponding formulation of the hypothesis using the matrix notation introduced before is

$$H_i:$$

$$R_E\boldsymbol{\theta} = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \theta_{control} \\ \theta_{traditional} \\ \theta_{individual} \\ \theta_{group} \\ \theta_{online} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = r_E,$$

$$R_I\boldsymbol{\theta} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_{control} \\ \theta_{traditional} \\ \theta_{individual} \\ \theta_{group} \\ \theta_{online} \end{bmatrix} > \begin{bmatrix} 0 \\ 0 \end{bmatrix} = r_I.$$

Note how each row of R_E and R_I matrices expresses an equality or an inequality constraint, respectively. For example, in the first row of R_E we have $\theta_{individual} - \theta_{group} = 0$ (i.e., $\theta_{individual} = \theta_{group}$) and in the first row of R_I we have $\theta_{traditional} - \theta_{control} > 0$ (i.e., $\theta_{traditional} > \theta_{control}$).

Now that we have understood how to define an informative hypothesis with equality and inequality constraints introducing an appropriate notation, let's see how to evaluate an informative hypothesis using the Bayes Factor with the encompassing prior approach.

6.2.2 Bayes Factor

The Bayes Factor of hypothesis H_1 against a competing hypothesis H_2 is defined as the ratio between the marginal likelihoods of the two hypotheses:

$$BF_{12} = \frac{Pr(Y|H_1)}{Pr(Y|H_2)} = \frac{\int l(Y|\theta_1, H_1)\pi(\theta_1|H_1)\, d\theta_1}{\int l(Y|\theta_2, H_2)\pi(\theta_2|H_2)\, d\theta_2}, \tag{6.2}$$

where Y indicates the data, θ_i is the vector of parameters under the hypothesis H_i, $l(Y|\theta_i, , H_i)$ is the likelihood function under the hypothesis H_i, and $\pi(\theta_i|H_i)$ is the prior of the parameters under the hypothesis H_i. The marginal likelihood $Pr(Y|H_i)$ can be interpreted as a measure of the plausibility of the data under H_i.

Therefore, the Bayes Factor BF_{12} quantifies the relative support of the data for the two competing hypotheses (and not the ratio between the probability of the two hypotheses). For values close to 1, the Bayes Factor indicates that H_1 and H_2 have similar support from the data. Larger values indicate evidence in favour of H_1, whereas values close to 0 indicate evidence in favour of H_2. The ratio between the posterior probabilities of the two hypotheses can be computed as

$$\frac{Pr(H_1|Y)}{Pr(H_2|Y)} = \overbrace{\frac{Pr(Y|H_1)}{Pr(Y|H_2)}}^{BF_{12}} \times \frac{Pr(H_1)}{Pr(H_2)}, \tag{6.3}$$

where $Pr(H_i)$ is the prior probability of H_i and $Pr(H_i|Y)$ is the posterior probability of H_i. For a detailed description of the Bayes Factor considering also its interpretation and application in different contexts see Heck et al. (2020), Mulder and Wagenmakers (2016), and Wagenmakers et al. (2010).

Note that to compute the marginal likelihood $Pr(Y|H_i)$ of a hypothesis H_i it is necessary to integrate the product between the likelihood and the prior. These integrals, however, are usually difficult to compute [in particular when hypotheses include order constraints]. In the case of hypotheses with equality and inequality constraints, however, it is possible to simplify the computation of the Bayes Factor using the encompassing prior approach.

Encompassing Prior Approach

The basic idea of the encompassing prior approach is to consider an informative hypothesis as a subset of the parameter space of an unconstrained model. Thus, to

evaluate the plausibility of a hypothesis we can consider the proportion of parameter space of the unconstrained model that satisfies the constraints. More specifically, given an informative hypothesis H_i and an unconstrained model H_u (or *encompassing model*) that does not contain any constraints on the parameters, if the prior under H_i is defined as a truncation of the proper prior under H_u (*encompassing prior*) according to the constraints, then the Bayes Factor between H_i and H_u can be written as

$$\begin{aligned} BF_{iu} &= \frac{Pr(\text{Inequality Const}|\text{Equality Const, Data, } H_u)}{\pi(\text{Inequality Const}|\text{Equality Const, } H_u)} \times \frac{Pr(\text{Equality Const}|\text{Data, } H_u)}{\pi(\text{Equality Const}|H_u)} \\ &= \frac{Pr(R_I\theta > r_I|R_E\theta = r_E, Y, H_u)}{\pi(R_I\theta > r_I|R_E\theta = r_E, H_u)} \times \frac{Pr(R_E\theta = r_E|Y, H_u)}{\pi(R_E\theta = r_E|H_u)}. \end{aligned} \tag{6.4}$$

The first term is the ratio between the conditional posterior probability and the conditional prior probability that the inequality constraints hold under the unconstrained model H_u given the equality constraints. The second term is the ratio between marginal posterior density and the marginal prior density of the equality constraints under H_u (the well-known Savage–Dickey density ratio; Dickey, 1971; Wetzels et al., 2010). In particular, the four elements can be interpreted as:

- The **conditional posterior probability** $Pr(R_I\theta > r_I|R_E\theta = r_E, Y, H_u)$ is a measure of the fit of the inequality constraints of H_i under H_u.
- The **conditional prior probability** $\pi(R_I\theta > r_I|R_E\theta = r_E, H_u)$ is a measure of the complexity of the inequality constraints of H_i under H_u.
- The **marginal posterior density** $Pr(R_E\theta = r_E|Y, H_u)$ is a measure of the fit of the equality constraints of H_i under H_u.
- The **marginal prior density** $\pi(R_E\theta = r_E|H_u)$ is a measure of the complexity of the equality constraints of H_i under H_u.

Summarizing at the denominators we have measures of the **complexity** of the inequality and equality constraints of the informative hypothesis H_i; at the numerators, instead, we have measures of the **fit** of the data to the inequality and equality constraints of the informative hypothesis H_i. Thus, if the hypothesis H_i, although applying constraints to the parametric space (less complexity), is still able to provide a good description of the data, the Bayes Factor will favour H_i. On the contrary, if the support of the data is poor, the Bayes Factor will favour the unconstrained model H_u.

The proof of the formulation of the Bayes Factor with the encompassing prior approach and the evaluation of its consistency (i.e., the probability of selecting the

correct hypothesis goes to 1 for the sample size going to infinity) is provided in Gu et al. (2018) and Mulder and Gelissen (2019). Note that slightly different notation is used here to enhance comprehension and underline that at the numerators we have values computed from the posterior of H_u, whereas at the denominators we have values computed from the prior of H_u.

To compute the Bayes Factor with the encompassing prior approach, only the prior and the posterior of the unconstrained model are required. To obtain them, first, we need to define the encompassing prior of the unconstrained model.

Definition of the Encompassing Prior

Prior specification is an important element as the resulting Bayes Factor value is affected by the prior choice. This aspect is particularly relevant in the case of equality constraints (Bartlett, 1957; Lindley, 1957), whereas inequality constraints are not affected as long as the prior is symmetric and centred to the focal point of interest (this aspect will be further discussed in Section 6.2.2).

To avoid arbitrary prior specification, different methods have been proposed in the literature. For example: Jeffreys-Zellner-Siow (JZS) objective priors do not require subjective specification (Bayarri & García-Donato, 2007; Jeffreys, 1961; Zellner & Siow, 1980); partial Bayes Factor (de Santis & Spezzaferri, 1999) defines the prior according to part of the data, whereas the remaining part is used to compute the Bayes Factor; intrinsic Bayes Factor (Berger & Pericchi, 1996) and fractional Bayes Factor (O'Hagan, 1995) are a variation of the partial Bayes Factor where priors are defined according to the average of all possible minimal subsets of the data or to a given small fraction of the data, respectively.

Both, Gu et al. (2018) and Mulder and Olsson-Collentine (2019) based their approach on the fractional Bayes Factor. Starting from a non-informative prior, a minimal fraction of the data is used to obtain a posterior that is subsequently used as proper prior, we refer to it as *fractional prior* (have you ever heard Lindley's 1972 famous quote *"Today's posterior is tomorrow's prior"*?). The remaining part of the data is used to compute the Bayes Factor. The two approaches, however, have an important difference. Gu et al. (2018) approximated the obtained fractional prior (as well as the posterior, see Section 6.2.2) to a (multivariate) normal distribution. In fact, due to large-sample theory (Gelman et al., 2013), parameter posterior distribution can be approximated to a (multivariate) normal distribution. On the contrary, Mulder and Olsson-Collentine (2019) provided an analytic solution in the case of linear regression models, obtaining an accurate quantification of the distribution.

All methods discussed above suggest objective procedures to avoid arbitrary prior specification. Nevertheless, using subjective (reasonable) priors according to previous information or experts' indications is still a possible approach. Keep

in mind, however, that prior specification (even if obtained through an *objective procedure*) affects the Bayes Factor results. Thus, it is fundamental to conduct a prior sensitivity analysis to evaluate the influence of prior specification on the results (Du et al., 2019; Schad et al., 2021).

Going back to our psychological treatment example, we could use independent normal distributions to specify the prior of each parameter of interest (i.e. $\theta_{control}$, $\theta_{traditional}$, $\theta_{individual}$, θ_{group}, and θ_{online}) obtaining as resulting prior a multivariate normal distribution with mean vector μ_θ and covariance matrix Σ_θ. In this way, we can still follow the approach proposed by Gu et al. (2018), based on normal approximation, that will simplify the computation of the Bayes Factor. For example, suppose that, according to experts' indications, a reasonable prior choice for all parameters of interest is a normal distribution with mean zero and standard deviation 2: $\mathcal{N}(0, 2)$. The resulting prior is a multivariate normal distribution with mean vector μ_θ and covariance matrix Σ_θ:

$$\pi(\boldsymbol{\theta}) \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$\text{where } \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_\theta = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}.$$

At this point, several authors underline the importance of centring the prior distribution on the constraints points of interest. Let's further discuss this issue in the next section.

Adjusting Prior Mean

When evaluating informative hypotheses with equality and inequality constraints using the Bayes Factor, the priors have to be centred on the focal points of interests (Jeffreys, 1961; Mulder, 2014; Zellner & Siow, 1980). In the case of equality constraints, centring the prior allows one to consider values close to the point of interest more likely a priori than distant values. This should be in line with researchers' expectations, otherwise one could question why testing that value.

In the case of inequality constraints, instead, this adjustment is done to guarantee that no constraint is favoured a priori. Consider the example represented in Figure 6.1 where the hypothesis $\theta_i > k$ is evaluated against $\theta_i < k$. Remember that, when computing the Bayes Factor, the prior probability that the constraint holds is used as a measure of the complexity of the hypothesis. Thus, if the prior is not centred on the focal point of interest (i.e., k), the less complex hypothesis (i.e.,

$\theta_i > k$) will be erroneously preferred a priori over the other (i.e., $\theta_i < k$). Only by centring the prior on the focal point, the hypotheses will be equally likely a priori.

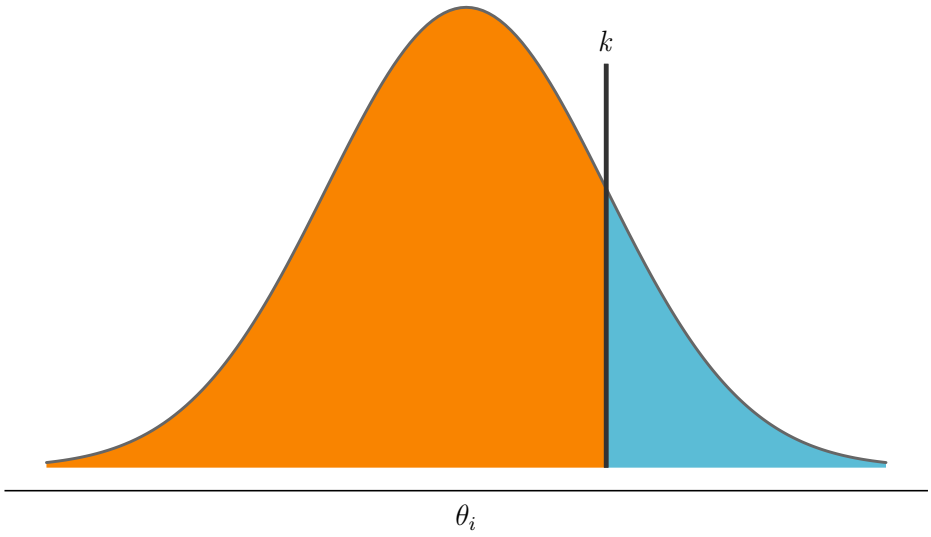

Figure 6.1: Example of non centered prior considering the constraints $\theta_i > k$ vs. $\theta_i < k$.

Centering the prior to the focal point, however, can be difficult in the case of complex hypotheses where constraints are defined as a linear combination of several parameters. To overcome this issue, Gu et al. (2018) proposed the following transformation of the parameters of interest:

$$\begin{gathered} \boldsymbol{\beta} = R\boldsymbol{\theta} - r \\ \text{with } \boldsymbol{\beta} = \begin{bmatrix} \beta_E \\ \beta_I \end{bmatrix}, \; R = \begin{bmatrix} R_E \\ R_I \end{bmatrix} \text{ and } r = \begin{bmatrix} r_E \\ r_I \end{bmatrix}, \end{gathered} \tag{6.5}$$

where $\boldsymbol{\theta}$ is the vector of original parameters, R is the matrix expressing equality and inequality constraints, and r is the vector with the constants of the equality and inequality constraints. Doing this, the informative hypothesis under evaluation becomes:

$$H_i : \quad \beta_E = 0 \quad \& \quad \beta_I > 0. \tag{6.6}$$

This parameter transformation has the advantage of simplifying the hypothesis expression without changing the original expectations. In fact, for example, evaluating $\theta_i > \theta_j$ is equivalent to evaluating $\beta_i = \theta_i - \theta_j > 0$. Thus, given the original

prior $\pi(\boldsymbol{\theta}) \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$, the prior of the new parameter vector $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}) \sim \mathcal{N}(\mu_\beta, \Sigma_\beta) = \mathcal{N}(R\boldsymbol{\theta} - r, R\Sigma_\theta R^T). \tag{6.7}$$

Note that this operation is nothing more than applying a linear transformation to the original multivariate normal distribution.[2] In order to do that, however, the matrix R must be *full-row-rank* (i.e., all rows are linearly independent). If this is not the case, the obtained matrix Σ_β will not be a proper covariance matrix. A possible solution is to follow the approach proposed by Mulder and Olsson-Collentine (2019) and Mulder (2016): a new matrix is defined selecting the maximum number of linearly independent rows and the remaining constraints are obtained as linear combinations (see Supplemental Material for more details https://claudiozandonella.github.io/Attachment/).

Now, to center the prior of $\boldsymbol{\beta}$ to the focal points, we can simply set the mean vector to zero. Thus, the adjusted prior of $\boldsymbol{\beta}$ is

$$\pi_{adj}(\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \Sigma_\beta) = \mathcal{N}(\mathbf{0}, R\Sigma_\theta R^T). \tag{6.8}$$

Considering the psychological treatment example, the obtained adjusted prior is $\mathcal{N}(\mathbf{0}, \Sigma_\beta)$ where:

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_\beta = R\Sigma_\theta R^T = \begin{bmatrix} 8 & -4 & 0 & 4 \\ -4 & 8 & 0 & 0 \\ 0 & 0 & 8 & -4 \\ 4 & 0 & -4 & 8 \end{bmatrix}.$$

So far we have defined the prior for the parameter vector $\boldsymbol{\theta}$ of the encompassing model. Moreover, we have obtained the adjusted prior for the new transformed parameter vector $\boldsymbol{\beta}$ that will allow us to properly evaluate the equality and inequality constraints. At this point, we need to compute the posterior of the encompassing model parameters.

Posterior Encompassing Model

Posterior distribution of the encompassing model parameters can be obtained through numerical approximation using Markov Chain Monte Carlo (MCMC) sampling algorithms, such as Metropolis–Hastings algorithm (Hastings, 1970) or Gibbs sampling (Geman & Geman, 1984). Bayesian statistical inference methods are implemented in all major statistical software. In R statistical software (R Core Team, 2021), for

[2] Given a multivariate normal distribution $Y \sim \mathcal{N}(\mu, \Sigma)$, the result of a linear transformation $AY + b$ is still a multivariate normal distribution with vector mean $\mu_t = AY + b$ and covariance matrix $\Sigma_t = A\Sigma A^T$.

example, the popular `brms` package (Bürkner, 2017, 2018), which is based on STAN (Stan Development Team, 2020), allows to easily conduct Bayesian inference.

Following Gu et al. (2018) approach, once we obtain the model posterior, we can approximate it to a (multivariate) normal distribution. Thus, the resulting posterior distribution is

$$Pr(\boldsymbol{\theta}|Y) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\Sigma}_\theta) \tag{6.9}$$

where posterior mean $\hat{\boldsymbol{\theta}}$ and posterior covariance $\hat{\Sigma}_\theta$ can be computed directly from the posterior draws. Next, we can obtain the posterior with respect to the vector parameters $\boldsymbol{\beta}$ applying the same transformation used for the prior distribution,

$$Pr(\boldsymbol{\beta}|Y) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma}_\beta) = \mathcal{N}(R\hat{\boldsymbol{\theta}} - r, R\hat{\Sigma}_\theta R^T). \tag{6.10}$$

At this point, we have both the prior and the posterior distributions of the parameters of interests of the encompassing model. Before proceeding, however, let's underline some important aspects.

When defining the prior, we adjusted the prior mean by centring it over the constraint focal points. This change would slightly influence the resulting posterior as well. However, as underlined by Gu et al. (2018), small prior changes will result in negligible changes on the posterior for large samples due to large-sample theory. For this reason, the authors do not adjust the posterior in their approach, but only the prior. Therefore, when computing the posterior we can simply consider the prior $\pi(\boldsymbol{\theta}) \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ defined at the beginning, without worrying about adjusting it.

In addition, note that, in Gu et al. (2018) original approach, the posterior mean $\hat{\boldsymbol{\theta}}$ and the posterior covariance $\hat{\Sigma}_\theta$ are not obtained from the posterior draws but they are computed directly from the sample data using the maximum likelihood estimate and the inverse of the Fisher information matrix, respectively. This has the advantage of being faster (posterior draws are not required) but it may be not possible for some complex models.

6.2.3 Computing the Bayes Factor

In the previous section we obtained the adjusted prior and the posterior of the vector of transformed parameters $\boldsymbol{\beta}$, respectively

$$\pi_{adj}(\boldsymbol{\beta}) \sim \mathcal{N}(0, \Sigma_\beta) \quad \text{and} \quad Pr(\boldsymbol{\beta}|Y) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\Sigma}_\beta).$$

Given the parameter transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$, we can rewrite the Formula 6.4 of the Bayes Factor between H_i and H_u as follow,

$$BF_{iu} = \frac{Pr(\beta_I > 0|\beta_E = 0, Y, H_u)}{\pi_{adj}(\beta_I > 0|\beta_E = 0, H_u)} \times \frac{Pr(\beta_E = 0|Y, H_u)}{\pi_{adj}(\beta_E = 0|H_u)}, \tag{6.11}$$

where β_I and β_E are defined in Formula 6.5 and represent the inequality and equality constraints, respectively.

Now, thanks to (multivariate) normal distribution approximation, we can easily compute the required conditional probabilities and marginal densities required to calculate the Bayes Factor.

Marginal Density

The marginal distribution of a subset of variables of a multivariate normal distribution is obtained simply discarding the variables to marginalize out. For example, given the adjusted prior of the psychological treatment example, $\pi_{adj}(\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \Sigma_\beta)$ where

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_\beta = \begin{bmatrix} 8 & -4 & 0 & 4 \\ -4 & 8 & 0 & 0 \\ 0 & 0 & 8 & -4 \\ 4 & 0 & -4 & 8 \end{bmatrix},$$

the marginal distribution of the equality constraints $\boldsymbol{\beta}_E$ is

$$\pi_{adj}(\boldsymbol{\beta}_E) \sim \mathcal{N}(\mu_{\beta_E}, \Sigma_{\beta_E}),$$

$$\text{where} \quad \mu_{\beta_E} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_{\beta_E} = \begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}.$$

At this point computing the density at $\beta_E = 0$ is elementary. In R, this can be done using the `dmvnorm()` function from the `mvtnorm` package (Genz et al., 2021). To compute the **marginal prior density** $\pi_{adj}(\beta_E = 0|H_u)$ of the psychological treatment example, we can use the following code:

```
# Prior info
mu_prior <- c(0, 0, 0, 0)
Sigma_prior <- matrix(c( 8,-4, 0, 4,
                        -4, 8, 0, 0,
                         0, 0, 8,-4,
                         4, 0,-4, 8), ncol = 4, byrow = TRUE)

# Marginal prior density at beta_1 = 0 and beta_2 = 0
mvtnorm::dmvnorm(x = c(0, 0),
                 mean = mu_prior[1:2],
                 sigma = Sigma_prior[1:2, 1:2])

## [1] 0.02297204
```

Analogously, it is possible to compute the **marginal posterior density** $\pi_{adj}(\beta_E = 0|Y, H_u)$ considering this time the estimated posterior mean vector $\hat{\boldsymbol{\beta}}$ and the estimated posterior covariance matrix $\hat{\Sigma}_\beta$.

Conditional Probability

To compute the conditional probability of the inequality constraints given the equality constraints in a multivariate normal distribution, we can use the `pcmvnorm()` function from the `condMVNorm` R-package (Varadhan, 2020). Considering the psychological treatment example, to calculate the **conditional prior probability** $\pi_{adj}(\beta_I > 0|\beta_E = 0, H_u)$ we can run the following code:

```
# Conditional prior probability that beta_3 > 0 and beta_4 > 0
# given beta_1 = 0 and beta_2 = 0
condMVNorm::pcmvnorm(
    lower = c(0, 0), upper = c(Inf, Inf), # inequality constraints
    mean = mu_prior, sigma = Sigma_prior,
    dependent.ind = 3:4,                  # inequality variables
    given.ind = 1:2, X.given = c(0, 0))   # equality vars and consts

## [1] 0.1451077
## attr(,"error")
## [1] 1e-15
## attr(,"msg")
## [1] "Normal Completion"
```

Analogously, it is possible to compute the **conditional posterior probability** $\pi_{adj}(\beta_I > 0|\beta_E = 0, Y, H_u)$ considering this time the estimated posterior mean vector $\hat{\boldsymbol{\beta}}$ and the estimated posterior covariance matrix $\hat{\Sigma}_\beta$.

At this point, we have all the elements required and we can easily compute the Bayes Factor BF_{iu}.

6.3 A Case Study: Hypotheses Testing in the Attachment Theory

In this section, we propose a real case study evaluating different informative hypotheses within the attachment theory. This will allow us to face the common issues found in real research applications. First, we provide the required background information regarding the attachment theory and the study aims and characteristics.

Subsequently, all the steps involved in the evaluation of informative hypotheses using the Bayes Factor with the encompassing prior approach are described. Finally, the results are briefly discussed considering the prior sensitivity analysis and limits of this approach.

All analyses were conducted using the R statistical software (v4.1.0; R Core Team, 2021). All materials, data, and analysis code are available at https://github.com/ClaudioZandonella/Attachment. The Supplemental Material with further details is also available online at https://claudiozandonella.github.io/Attachment/.

6.3.1 Background Information

The Attachment Theory

The attachment theory originates from the pioneering work of Bowlby (1969) and Ainsworth (1970). They postulates that children in stressful situations actively seek proximity of the caregiver through some attachment behaviors (e.g., crying; moving towards the caregiver) in order to fulfil the evolutionary goal of protection from dangers. The main tenet of attachment theory is that the relationships with the caregivers that children develop in the early stages of their life (i.e., *attachment bond*) will affect children social and emotional future development (Cassidy & Shaver, 2016). Besides behavior, people construct mental representations, or working models, of the self and significant others based on their interpersonal experiences. Four main attachment styles have been recognized in the literature according to different internal representations:

- **Secure Attachment** - children who are securely attached display optimal emotional regulation and they consider the caregiver as a secure base.
- **Anxious Attachment** - anxious children manifest high levels of anxiety in stressful situations and their relationships with the caregivers is ambivalent displaying anger or helplessness.
- **Avoidant Attachment** - avoidant children mask distress in stressful situations displaying little emotions and their relationships with the caregivers is characterized by little involvement.
- **Fearful Attachment** - fearful children lack adequate emotional regulation in stressful situations with the risk of displaying non organized behaviors.

Attachment theory is one of the main and most supported theories in psychology (Cassidy & Shaver, 2016). In the literature, however, there is still an open debate on the relative role of mother and father attachment on children's social-emotional development. Four different main theoretical perspectives have been identified (Bretherton, 2010):

- **Monotropy Theory** - only the principal attachment figure (usually the mother) has an impact on children's development.
- **Hierarchy Theory** - the principal attachment figure (usually the mother) has a greater impact on children development than the subsidiary attachment figure (usually the father).
- **Independence Theory** - all attachment figures are equally important but they affect the children's development differently.
- **Integration Theory** - to understand the impact on children's development it is necessary to consider all attachment relationships taken together.

Contrasting results have been reported by studies investigating which is the "*correct*" theory. No study, however, has tried to properly evaluate the different theoretical perspectives by directly comparing the different hypotheses.

Present Study

The present study aims to directly compare the four different theoretical perspectives regarding the role of father and mother attachment, using the Bayes Factor with the encompassing prior approach.

In the analysis, $n = 847$ Italian children (50.65% Females) between 8 and 12 years old (*middle childhood*, third to sixth school grade) were included. Attachment towards the mother and the father was measured separately using the italian version of the Experiences in Close Relationships Scale - Revised Child version (ECR-RC; Brenning et al., 2014; Marci et al., 2019) completed by the children. Subsequently, two separate cluster analyses (i.e., one for mother scores, one for father scores) were performed. Both analyses supported the existence of the four attachment profiles (see above). The results of the classification are reported in Table 6.1.

Table 6.1: Attachment styles frequencies ($n_{subj} = 847$).

	Father Attachment				
Mother Attachment	**Secure**	**Anxious**	**Avoidant**	**Fearful**	**Total**
Secure	125	49	49	8	231
Anxious	51	100	98	37	286
Avoidant	25	67	126	12	230
Fearful	5	14	38	43	100
Total	206	230	311	100	847

Children's social-emotional development was measured using the Strength & Difficulties Questionnaire (SDQ; A. Goodman et al., 2010; R. Goodman, 1999) completed by the teachers. Separate scores for externalizing and internalizing problems were obtained as sum of the questionnaire items. In the analysis, however, only externalizing problems are considered as teachers are expected to be better at reporting externalizing problems than internalizing problems. The distribution of externalizing problems (Mean = 3.35; SD = 3.91; Median = 2.0) is presented in Figure 6.2.

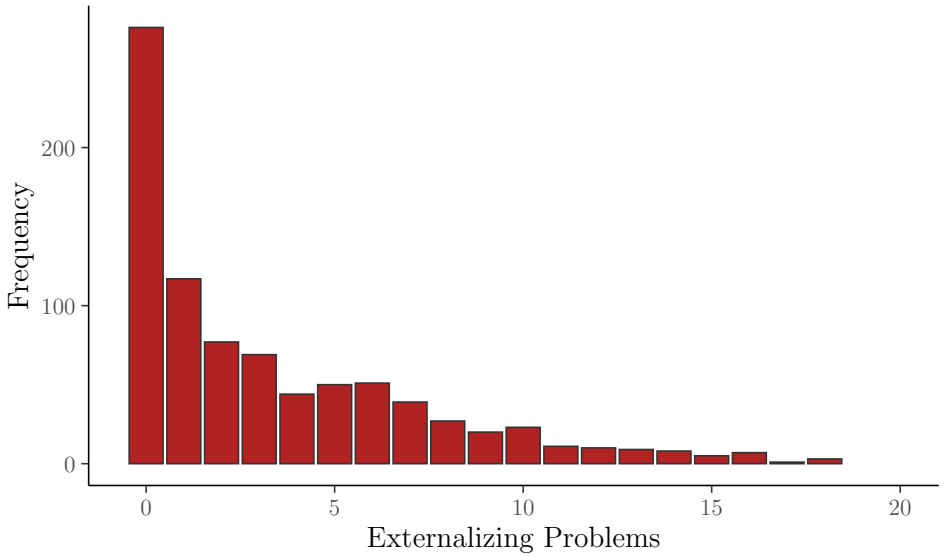

Figure 6.2: Distribution of externalizing problems ($n_{subj} = 847$).

Externalizing problems according to attachment styles are reported in Table 6.2. More information about the sample, descriptive statistics, cluster analysis, and analysis of internalizing problems can be found in the Supplemental Material available online https://claudiozandonella.github.io/Attachment/.

Table 6.2: Externalizing problems according to attachment styles ($n_{subj} = 847$).

	Father Attachment							
	Secure		Anxious		Avoidant		Fearful	
Mother Attachment	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Secure	2.63 (3.57)	1.0	3.45 (4.48)	2.0	1.61 (2.13)	1.0	2.88 (3.44)	2.0
Anxious	3.69 (4.07)	2.0	3.01 (3.61)	2.0	3.32 (4.12)	2.0	4.05 (3.61)	3.0
Avoidant	2.84 (3.34)	1.0	3.31 (3.65)	2.0	3.71 (4.19)	2.0	3.75 (4.81)	1.0
Fearful	7.60 (4.04)	8.0	4.64 (3.84)	4.5	4.76 (4.67)	3.0	4.26 (4.07)	4.0

6.3.2 Evaluating Hypotheses with Bayes Factor

Formalization of Informative Hypotheses

Each theoretical perspective was formalized into a different informative hypothesis taking into account its own theoretical tenets, main evidence in the literature, and authors' clinical experience in the field.

The following notation is used to formalize the hypothesis. M and F are used to indicate the attachment towards the mother and the father, respectively. The specific attachment style is specified in the subscript where S indicates secure attachment, Ax anxious attachment, Av avoidant attachment, and F fearful attachment. For example, F_{Av} represents children with avoidant attachment towards the father.

Note that when we do not expect interaction between mother and father attachment, we can consider the role of the two parents separately. Whereas, if an interaction is expected, it is necessary to take into account the unique combination of mother and father attachment. Thus, for example, we use $M_S F_{Ax}$ to indicate children with secure attachment towards the mother and anxious attachment towards the father. Moreover, $*$ subscript is used to indicate any attachment style and a set of subscripts is used to indicate "*one among*". For example, $M_S F_{Ax;Av}$ represents children with secure attachment towards the mother and anxious or avoidant attachment towards the father.

To correctly interpret the following Figures, note that the order is important, whereas the actual values are only indicative.

Null Hypothesis. This is a reference hypothesis where mother attachment and father attachment are expected to have no effect and only gender differences are taken into account (see Section 6.3.2 for the actual model definition). The hypothesis is represented in Figure 6.3:

$$M_* = 0,$$
$$F_* = 0.$$

Monotropy Hypothesis. Father attachment is expected to have no effect, whereas considering mother attachment we expect the following order: secure children with the lowest level of problems, anxious and avoidant children with similar levels of problems, fearful children with the highest levels of problems. The hypothesis is represented in Figure 6.4:

$$M_S < M_{Ax} = M_{Av} < M_F,$$
$$F_* = 0.$$

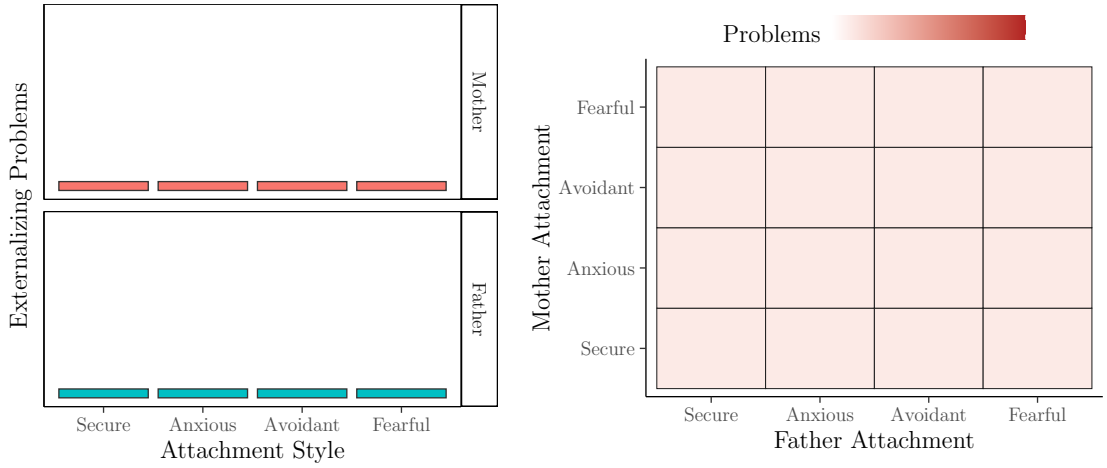

Figure 6.3: Null Hypothesis. Expected externalizing problems according to mother and father attachment.

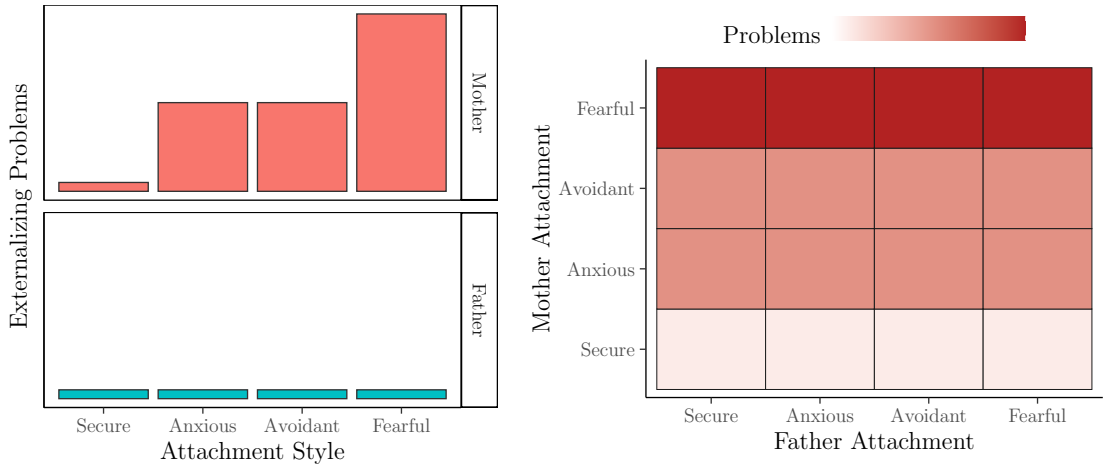

Figure 6.4: Monotropy Hypothesis. Expected externalizing problems according to mother and father attachment.

Hierarchy Hypothesis. Father attachment is expected to follow the same pattern as the mother attachment, but its influence is expected to be smaller. The hypothesis is represented in Figure 6.5:

$$M_S < M_{Ax} = M_{Av} < M_F,$$
$$F_S < F_{Ax} = F_{Av} < F_F,$$
$$F_{Ax} < M_{Ax}; \quad F_{Av} < M_{Av}; \quad F_F < M_F.$$

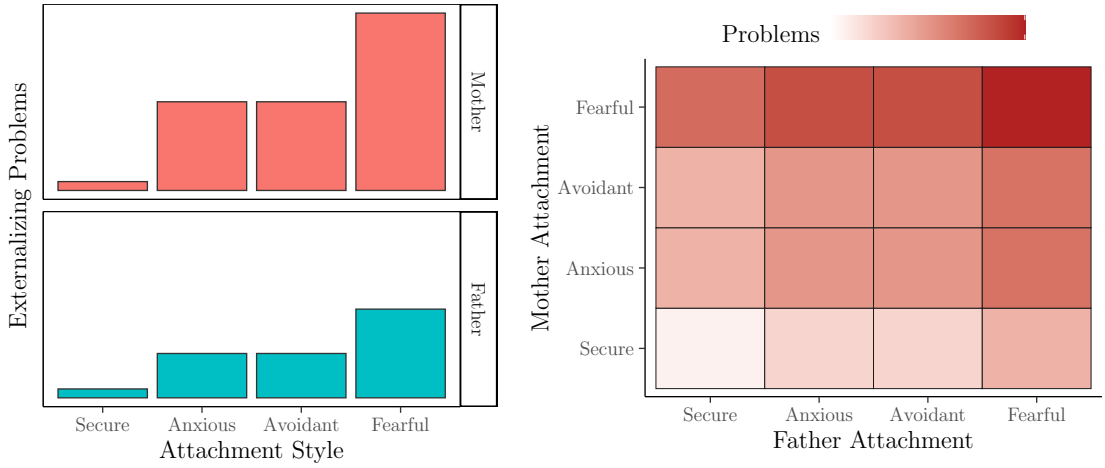

Figure 6.5: Hierarchy Hypothesis. Expected externalizing problems according to mother and father attachment.

Independence Hypothesis. Mother and father attachment are expected to affect children outcomes differently. In this case, we considered avoidant attachment towards the father as a condition of higher risk. The hypothesis is represented in Figure 6.6:

$$M_S < M_{Ax} = M_{Av} < M_F,$$
$$F_S < F_{Ax} < F_{Av} < F_F.$$

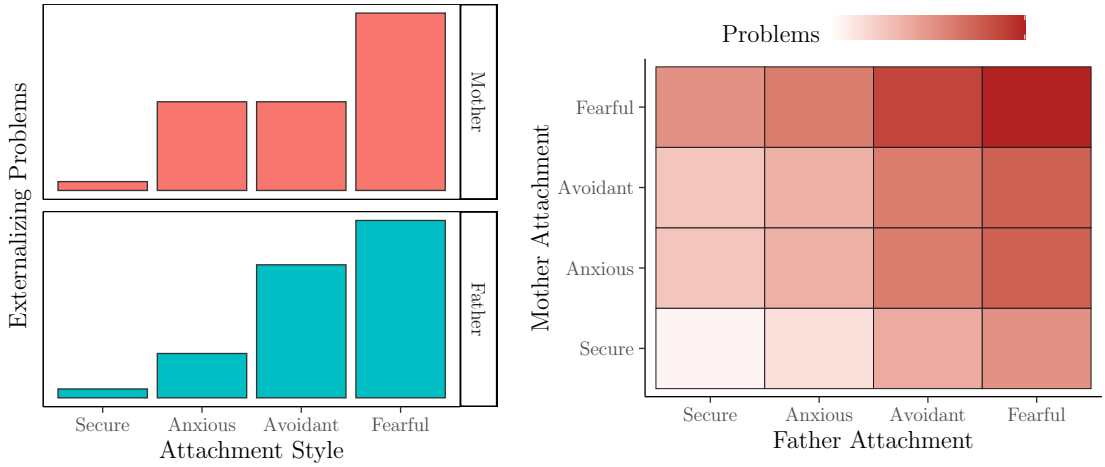

Figure 6.6: Independence Hypothesis. Expected externalizing problems according to mother and father attachment.

Integration Hypothesis. Mother and father attachment are expected to interact. In this case, we consider secure attachment as a protective factor and fearful attachment as a risk condition. We do not specify the conditions $M_S F_F$ and $M_F F_S$ as their frequency is very low (1.5% of the sample). The hypothesis is represented in Figure 6.7:

$$M_S F_S < \{M_S F_{Ax;Av} = M_{Ax;Av} F_S\} < M_{Ax;Av} F_{Ax;Av} < \{M_F F_{Ax;Av} = M_{Ax;Av} F_F\} < M_F F_F,$$

with

$$M_{Ax} F_{Ax} = M_{Ax} F_{Av} = M_{Av} F_{Ax} = M_{Av} F_{Av}.$$

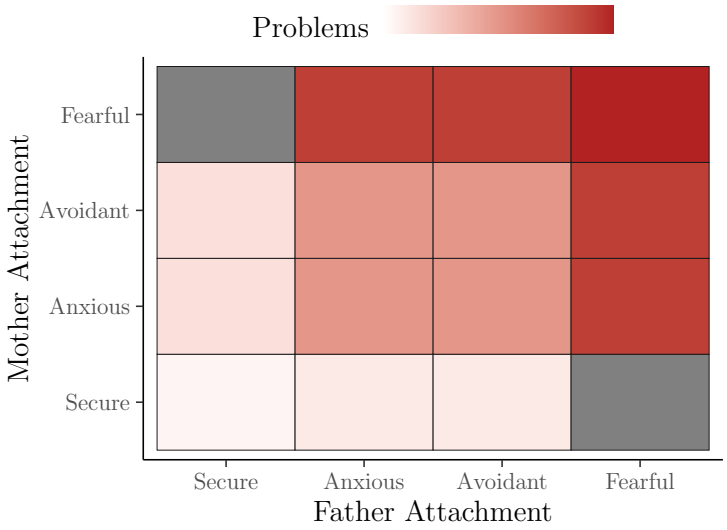

Figure 6.7: Integration Hypothesis. Expected externalizing problems according to mother and father attachment.

Definition of the Encompassing Model

A Zero-Inflated Negative Binomial (ZINB) mixed effect model is defined to take into account the characteristics of the dependent variable and its distribution (see Figure 6.2 and see Supplemental Material for more details regarding the analysis of zero inflation https://claudiozandonella.github.io/Attachment/):

$$y_{ij} \sim ZINB(p_{ij}, \mu_{ij}, \phi), \tag{6.12}$$

where p_{ij} is the probability of an observation y_{ij} being an extra zero (i.e., a zero not coming from the Negative Binomial distribution) and $1-p_{ij}$ indicates the probability

of a given observation y_{ij} being generated form a Negative Binomial distribution with mean μ_{ij} and variance $\sigma^2_{ij} = \mu_{ij} + \frac{\mu^2_{ij}}{\phi}$. Moreover, we define

$$\begin{aligned} p_{ij} &= \text{logit}^{-1}(X_i^T \beta_p + Z_j^T u_p), \\ \mu_{ij} &= \exp(X_i^T \beta_\mu + Z_j^T u_\mu). \end{aligned} \tag{6.13}$$

That is, both $\boldsymbol{p}$ and $\boldsymbol{\mu}$ are modelled separately according to fixed and random effects. In particular, we consider the children's classroom ID as a random effect in both cases to account for teachers' different ability to evaluate children's problems. While, regarding fixed effects, only the role of gender is considered for $\boldsymbol{p}$, whereas for $\boldsymbol{\mu}$ the interaction between mother and father attachment are included together with gender. In the R formula syntax, we have

```
# Regression on p
p ~ gender + (1|ID_class)

# Regression on mu
mu ~ gender + mother * father + (1|ID_class)
```

The parameters of interest (i.e., those related to mother and father attachment interaction) and the parameter related to gender differences are unbounded. Thus, we can simply specify a normal distribution with mean 0 and standard deviation of 3, $\mathcal{N}(0, 3)$, as reasonable prior. This prior is intended to be non-informative but without being excessively diffuse[3]. The influence of prior specification is subsequently evaluated in a prior sensitivity analysis. Regarding the other nuisance parameters (i.e., intercepts, random effects and shapes parameters) `brms` default priors are maintained (see Supplemental Material for further details https://claudiozandonella.github.io/Attachment/). The encompassing model was estimated using 6 independent chains with 10,000 iterations (warm-up 2,000).

Hypothesis Matrices

Before computing the hypothesis matrix for each informative hypothesis, it is important to consider the contrasts coding and the resulting parametrization of the encompassing model. For mother and father attachment, default treatment contrasts are used (Schad et al., 2020) considering secure attachment as the reference

[3]Considering 1 as intercept (note that $exp(1)$ is approximately the sample mean value), values included within one standard deviation, $exp(1 \pm 1 \times SD)$, range between 0 and 55. Although externalizing problems are bounded between 0 and 20, prior predicted values are still reasonable as they cover all possible values without including excessively large values. More diffuse priors would result in values with a higher order of magnitude and tighter priors would exclude plausible values (see Supplemental Material for further details https://claudiozandonella.github.io/Attachment/)

category. Therefore, model intercept represents children with secure attachment towards both parents ($M_S F_S$) and we have parameters indicating the main effects of mother and father attachment and other parameters for the interaction effect.

Now, given the informative hypotheses and the parametrization of the encompassing model, we can obtain the respective hypotheses matrices. To do that, first, we need the model matrix with all the conditions of interest. Note that we have to consider only conditions relevant to the constraints (i.e., those related to mother and father attachment) ignoring other nuisance conditions (i.e., gender and classroom ID).

Subsequently, we can derive the required equality and inequality constraints. In particular, with hypotheses that do not expect interaction between mother and father attachment (i.e., monotropy, hierarchy, and independence hypotheses), all interaction terms are set equal to zero and main effects are obtain considered the reference level of the other parent (i.e., $M_{Ax} F_S$ is the main effect of anxious attachment towards the mother). As an example, consider the following code:

```
# Define relevant conditions
attachment <- c("S", "Ax", "Av", "F")
new_data <- expand.grid(
  mother = factor(attachment, levels = attachment),
  father = factor(attachment, levels = attachment)
)

# Get model matrix (removing intercept)
mm <- model.matrix(~ mother * father, data = new_data)[,-1]
rownames(mm) <- paste0("M_",new_data$mother, "_F_", new_data$father)

# Get constraints main effect
# M_Ax = M_Av  --->  M_Ax_F_S - M_Av_F_S = 0
# F_F > F_Av   --->  M_S_F_F - M_S_F_Av > 0
rbind(mm["M_Ax_F_S", ] - mm["M_Av_F_S", ],
      mm["M_S_F_F", ] - mm["M_S_F_Av", ])

# Get constraints interaction
# M_F_F_Ax = M_Ax_F_F  --->  M_F_F_Ax - M_Ax_F_F = 0
# M_Ax_F_Av > M_S_F_Ax  --->  M_Ax_F_Av - M_S_F_Ax > 0
rbind(mm["M_F_F_Ax", ] - mm["M_Ax_F_F", ],
      mm["M_Ax_F_Av", ] - mm["M_S_F_Ax", ])
```

In this way, we can easily specify all the constraints obtaining the respective hypothesis matrix (R) and the vector with the constraints constant values (r) for

each hypothesis. In all our hypotheses, however, no constraints include constant values. Thus, r is always a vector of zeros and it can be ignored. Moreover, also the intercept is ignored because there are no constraints that include $M_S F_S$ alone (e.g., $M_S F_S > k$) but is always compared with other groups (e.g., $M_S F_S < M_{Ax} F_S$). Thus, the inclusion of the intercept is redundant and we can ignore it.

Full hypothesis matrix specification for each hypothesis is available in the Supplemental Material https://claudiozandonella.github.io/Attachment/.

Computing the Bayes Factor

So far we defined the hypotheses matrices, specified the encompassing prior, and obtained the model posterior distribution. To compute the Bayes Factor, we now need the adjusted prior and the posterior of the transformed parameters vector $\boldsymbol{\beta}$ (i.e., the parameters that identify the constraints) for each hypothesis.

- **Adjusted Prior $\boldsymbol{\beta}$.** As reported in Section 6.2.2, adjusted prior is required to properly evaluate the constraints. Applying the Equation 6.7, we obtain the prior for the transformed parameter and then we set the mean vector to zero.

- **Posterior $\boldsymbol{\beta}$.** The same transformation used for the prior can be applied, this time considering the estimated posterior mean $\hat{\boldsymbol{\theta}}$ and the estimated posterior covariance $\hat{\Sigma}_\theta$, to obtain the posterior distribution of the transformed parameters vector $\boldsymbol{\beta}$.

Note that equation 6.7 requires the hypothesis matrix R to be *full-row-rank* (i.e., all constraints are linearly independent). However, this is not the case of the hierarchy hypothesis. To overcome this limit, we can use the solution proposed by Mulder and Olsson-Collentine (2019) and Mulder (2016): a new matrix is defined selecting the maximum number of linearly independent rows and the remaining constraints are obtained as linear combinations. Detailed information is available in the Supplemental Material https://claudiozandonella.github.io/Attachment/.

Now, we have all the elements required to compute the Bayes Factor for each hypothesis as described in Section 6.2.3. Moreover, assuming that each hypothesis is equally likely a priori, we can calculate the posterior probability of each hypothesis as (Gu et al., 2014; Hoijtink, 2012),

$$\text{Posterior Probability} H_i = \frac{BF_{iu}}{\sum_i BF_{iu}}. \tag{6.14}$$

6.3.3 Results and Prior Sensitivity Analysis

Bayes Factor and posterior probability of each hypothesis are reported in Table 6.3. Results clearly indicate that, among the considered hypotheses, the Monotropy Hypothesis is the most supported by the data.

Table 6.3: Bayes Factor encompassing model and hypothesis posterior probabilities ($n_{subj} = 847$).

Hypothesis	Bayes Factor	Posterior Probability	
Null	2.9e+11	0.01	
Monotropy	2.6e+13	0.98	
Hierarchy	2.7e+11	0.01	
Independence	3.9e+09	0.00	
Integration	3.2e+09	0.00	

Remember, however, that prior specification affects the Bayes Factor results. It is recommended, therefore, to evaluate also the results obtained using different prior settings. In particular, we considered as possible priors for the parameters of interest:

- $\mathcal{N}(0, .5)$ - unreasonable tight prior
- $\mathcal{N}(0, 1)$ - tighter prior
- $\mathcal{N}(0, 3)$ - original prior
- $\mathcal{N}(0, 5)$ - more diffuse prior
- $\mathcal{N}(0, 10)$ - unreasonably diffuse prior

The results of the prior sensitivity analysis are reported in Table 6.4.

Overall results consistently indicate the Monotropy Hypothesis as the most supported by the data. However, we can observe two distinct patterns. As the prior gets more diffuse, the order of magnitude of the Bayes Factor comparing each hypothesis with the encompassing model increases. Moreover, the probability of the Null Hypothesis increases with more diffuse prior, whereas the probabilities of the Hierarchy, Independence and Integration Hypothesis increases with tighter priors.

Table 6.4: Bayes Factor encompassing model v and hypothesis posterior probabilities (PP) under different prior settings ($n_{subj} = 847$).

	$\mathcal{N}(0, .5)$		$\mathcal{N}(0, 1)$		$\mathcal{N}(0, 3)$		$\mathcal{N}(0, 5)$		$\mathcal{N}(0, 10)$	
Hypothesis	**BF**	**PP**	**BF**	**PP**	**BF**	**PP**	**BF**	**PP**	**BF**	**PP**
Null	8.2e+01	0.00	9.4e+04	0.00	2.9e+11	0.01	4.7e+14	0.03	1.5e+19	0.11
Monotropy	1.2e+05	0.67	6.3e+07	0.90	2.6e+13	0.98	1.6e+16	0.97	1.2e+20	0.89
Hierarchy	4.9e+04	0.28	6.1e+06	0.09	2.7e+11	0.01	6.7e+13	0.00	1.2e+17	0.00
Independence	4.4e+03	0.02	2.6e+05	0.00	3.9e+09	0.00	5.8e+11	0.00	5.1e+14	0.00
Integration	4.6e+03	0.03	3.3e+05	0.00	3.2e+09	0.00	3.3e+11	0.00	1.5e+14	0.00

To interpret these patterns, remember that order constraints are insensitive to the distribution specification as long as the distribution is symmetric and centred on the constraint focal point. On the contrary, equality constraints are highly affected by the prior definition. As the prior gets more diffuse, the density value at zero decreases as well and, even for posterior distributions centred far from zero, the densities ratio at zero will favour the posterior. Thus, as the prior gets more diffuse, the Bayes Factor will favour more and more the hypotheses with equality constraints, whereas for tighter prior the Bayes Factor will strongly penalize hypotheses with equality constraints if these are not correct (see Figure 6.8).

All the defined hypotheses include equality constraints. Thus, for more diffuse prior we observe that the order of magnitude of the Bayes Factor comparing each hypothesis with the encompassing model increases. Moreover, the hypothesis with a higher number of equality constraints (e.g., Null Hypothesis) will be favoured over hypotheses with a smaller number of equality constraints (e.g., Hierarchy, Independence and Integration Hypothesis).

Limits

When interpreting the results, it is important to take into account the limits of the Bayes Factor. First of all, the selected hypothesis is relative to the data and the set of hypotheses considered. That means that we should not interpret the selected hypothesis as the only "*correct*" one. Maybe we did not consider some important aspects and new hypotheses may reveal themselves to be actually much better than the previous ones. For example, regarding the attachment results, taking into account the interaction between children's gender and attachment may be fundamental (do you expect that attachment towards the father plays a different role between girls and boys?). This limit, however, becomes an advantage as it forces the researchers to focus on the definition of the hypotheses: to obtain good answers first we need to ask the right questions.

Moreover, considering a unique single winning hypothesis may be limiting as

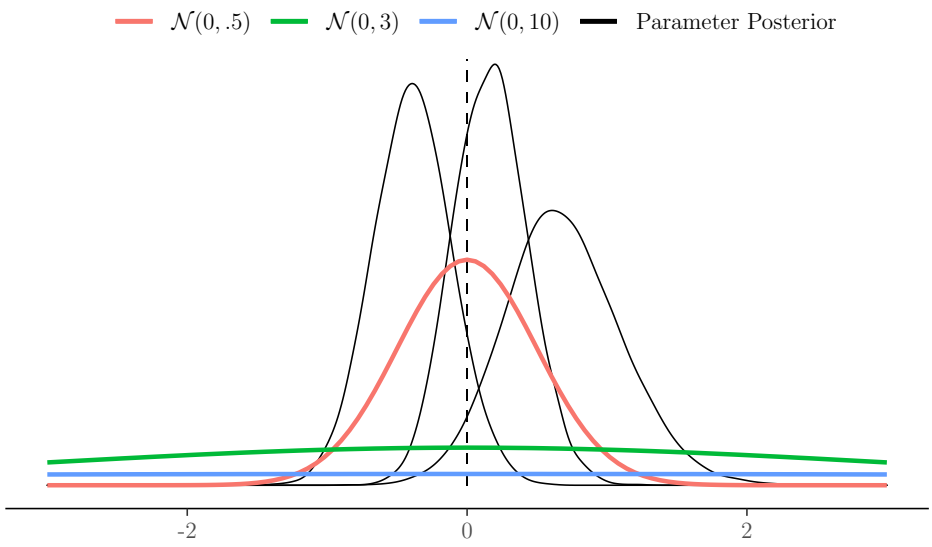

Figure 6.8: Evaluating densities at 0 for different prior settings and a selection of parameters posteriors.

different hypotheses could help to explain different aspects of the process under investigation. For example, mother and father attachment may play a different role at different ages. Thus, the selected hypothesis may vary according to children's age or other conditions. This should warn against interpreting the result in terms of winning/losing focusing uniquely on the selected hypothesis and discarding other hypotheses from further investigations.

Another important limit we observed is the sensitivity of the results to the prior specification. Gelman et al. (2013) do not recommend the use of the Bayes Factor given its sensitivity to model definition arbitrary choices. Evaluating the results under different conditions and transparently discussing the reasons behind arbitrary choices would help to interpret results more consciously (Schad et al., 2021).

Finally, another limit related to the use of the Bayes Factor with the encompassing prior approach is that we do not obtain the actual estimates of the parameters posterior. The only information we get is the selected hypothesis, but we have no other information regarding the actual parameters values. Therefore, for example, we are not able to assess the actual effect sizes. To overcome this limit we should obtain via Bayesian inference the actual posterior of the model parameters given the prior, formalized according to parameter constraints, and the observed data (for an introduction to Bayesian estimation see J. K. Kruschke & Liddell, 2018b). However, posterior estimation under inequality constraints requires the use of ap-

propriate advanced computational algorithms (see Ghosal, 2022; Kato & Hoijtink, 2006).

6.4 Conclusion

In this paper, we presented a detailed description and an applied example of the Bayes Factor with the encompassing prior approach. As discussed in the introduction, this approach has several advantages over the traditional Null Hypothesis Significance Testing (NHST): it is possible to evaluate complex hypotheses with equality and inequality constraints; multiple hypotheses can be tested simultaneously; Bayes Factor and Posterior Probabilities allow to quantify the relative evidence form the data in favour of the hypotheses.

This approach is of particular interest to the researchers, because informative hypotheses allow them to formalize expectation with great flexibility. In this paper we considered only the presence/absence of an effect (i.e., $\theta = 0$) or the expected effects order (i.e., $\theta_i > \theta_j$). However, in case of clear indications about the actual effects, researchers can specify the expected range of values for the effects of interest (i.e., $a < \theta_i < b$). This allows even a greater level of precision in the formalization of informative hypotheses.

As underlined by Gu et al. (2018, p.229), "this class of informative hypotheses covers a much broader range of scientific expectations than the class of standard null hypotheses. In addition, by testing competing informative hypotheses directly against each other a researcher obtains a direct answer as to which scientific theory is most supported by the data".

The available literature, however, is often technical and difficult to follow for non-experts in this approach. This paper filled this gap by providing a detailed description of all steps and elements involved in the computation of the Bayes Factor with the encompassing prior approach. This will enhance researchers' awareness regarding all pros and cons of this method, avoiding applying statistical techniques as black boxes.

Round Table

1. I would have liked to see the application of the Design Analysis covered in the first part also to the studies presented in the second part.

   Answer: Actually, we conducted a retrospective design analysis for the study regarding the stereotype threat (Chapter 5). However, we did not include it in the thesis to avoid the chapter being too lengthy. The retrospective design analysis can be found in the Supplemental Material of the original main article (https://osf.io/3u2jd/; see Section 7.2 Power analysis). Considering the attachment

study (Chapter 6), instead, no retrospective design analysis was conducted for two main reasons. First, defining appropriate plausible effect sizes for each effect of interest is difficult as the number of parameters and model complexity increase. This is one of the limits of design analysis and power analysis. Second, design analysis (as well as power analysis) is developed within the frequentist statistical approach. The definition of power in the Bayesian approach and the evaluation of frequentist properties of the Bayesian methods are currently discussed in the literature. Thus, the extension of the design analysis to the Bayesian approach is surely an interesting future step.

7
Discussion

We did it! Finally, we are at the end of this thesis. It was a really long journey into statistical inference that leads us to reconsider how we apply statistical methods to answer our research questions.

We started from the Null Hypothesis Significance Testing (NHST), the dominant statistical approach in Psychology and Social Sciences. The NHST approach is not inherently wrong per se, but its misuse and misinterpretation, in what Gigerenzer et al. (2004) defined as the "*Null Ritual*", is considered as one of the causes of the ongoing replicability crisis. In fact, when selecting for significance in underpowered studies evaluating complex multivariate phenomena with noisy data (all very common conditions in psychology), it is really easy to obtain misleading and unreliable results.

To evaluate the inferential risks related to effect size estimation when selecting for significance, we presented, in the first part of the thesis, the Design Analysis framework. In particular, in Chapter 2 we introduced the elements of the design analysis illustrating its advantages over traditional power analysis considering Cohen's d as a measure of effect size. In Chapter 3, we extended design analysis to the case of Pearson's correlation coefficients. Finally, in Chapter 4, we presented the `PRDA` R-package that allows researchers to perform prospective and retrospective design analysis in the case of Pearson's correlation between two variables or mean comparisons.

Even though the design analysis framework allowed us to evaluate the inferential risks related to effect size estimation, the main drawback of the NHST approach still remains: the NHST does not answer the question researchers are usually interested

in. Using the NHST approach, researchers can not evaluate the plausibility of their hypotheses, but only the evidence against the null hypothesis. To properly evaluate research hypotheses, in the second part of the thesis, we moved away from the NHST towards the model comparison approach.

Model comparison allows us to evaluate the relative evidence in favour of one hypothesis according to the data. In Chapter 5, we introduced the model comparison approach using the information criteria to select the preferred models. Information criteria assess models predictive ability penalizing for model complexity. However, they are not suitable for comparing informative hypotheses with inequality constraints. Thus, in Chapter 6, we introduced a different approach based on the Bayes Factor with encompassing prior. This approach allows researchers to easily evaluate informative hypotheses with equality and inequality constraints on the model parameters.

The aim of this thesis, however, was not only to present different approaches for testing research hypotheses but also to enhance researchers' statistical reasoning, one of the main ingredients for correct inference. During this journey into statistical inference, we stressed the importance of choosing appropriate statistical techniques, rather than applying statistical methods as black boxes. As pointed out by Gigerenzer and Marewski (2015, p. 422), *"if statisticians agree on one thing, it is that scientific inference should not be made mechanically."* In this regard, we think that focusing on modeling rather than testing enhances researchers' statistical reasoning. In fact, modeling forces researchers to consider the data generative process of the phenomena of interest and formalize their research hypotheses.

We could consider the mechanical application of the NHST approach just as an inappropriate tradition inherited from the early stages of research in Psychology when researchers were simply evaluating if "*there is something going on*". However, once the possible elements involved in the phenomena under study are identified, the NHST approach does not allow us to understand "*what actually is going on*". To do that, we introduce the model comparison approach using the information criteria or the Bayes Factor to evaluate informative hypotheses.

So, are these the solution to all our problems? Well, of course not. These are just the first steps towards a more conscious way to conduct statistical inference in research. However, many other steps can be done to improve statistical inference. For example, considering "*Causal Inference*". If you like feeling lost and questioning everything we have learned so far, please see McElreath (2021a). In a series of three posts[1], McElreath made me doubt if everything I did in three years was just a "*Causal Salad*" and I fell disappointed that I have never heard before about "*Causal Inference*". But, to go into this, another PhD would be required.

[1]On YouTube, there is also one of McElreath's (2021b) wonderful lectures about "*Causal Inference*": https://www.youtube.com/watch?v=KNPYUVmY3NM

7.1 Collaborating for Better Research

Finally, I want to underline how the replicability crisis has been also a great opportunity to improve research in psychology and social sciences. The replicability crisis created lots of concern in the literature. At the same time, however, it allowed rethinking traditional practices moving towards new approaches to improve research quality. In particular, the *Open Science Movement* spread across all disciplines fostering transparency and accessibility of scientific results and enhancing awareness about the social role of scientific research. The strong impact of this movement is changing the rules, these are exciting times.

Nowadays, transparency and reproducibility have become fundamental requirements for high-quality research. In this regard, domain knowledge, expertise in statistics, and programming skills[2], all are indispensable elements of modern research. This often requires the collaboration between researchers with different backgrounds, working together to reach the same goal: doing better research. Collaboration, however, might be difficult to achieve given the different competencies and skills. To facilitate this, we introduce in Appendix A `trackdown`, an R package offering a simple solution for collaborative writing and editing of reproducible documents (i.e. R-Markdown documents). Our small contribution to a gReat community!

> *If all you have is a hammer, everything looks like a nail*
>
> — **Maslow (1966)**

> *What makes science so powerful is that it's self-correcting*
>
> — **Anonymous**

[2]Are you looking for another amazing McElreath's (2020a) talk? *Science as Amateur Software Development* https://www.youtube.com/watch?v=zwRdO9_GGhY

A

`trackdown`: An R Package for Enhancing Collaborative Writing[1]

A.1 Introduction

Literate programming allows combining narrative text and computer code to produce elegant, high quality, and reproducible documents. These are fundamental requirements of modern open science workflows fostering transparency and reproducibility of scientific results.

A major downside of literate programming, however, is the lack of tools for collaborative writing and editing. Producing a document following a literate programming approach requires programming skills which can complicate the collaboration with colleagues who lack those skills. Furthermore, while commonly used version control systems (e.g., Git) are extremely powerful for collaborating on the writing of computer code, they are less efficient and lack the interactivity needed for collaborating on the writing of the narrative part of a document. On the contrary, common word processors (e.g., Microsoft Word or Google Docs) offer a smoother experience in terms of real-time editing and reviewing.

[1]This chapter is adapted from the `trackdown` package documentation (Kothe et al., 2021), in which I contributed to conceiving the original idea, developing the package, and writing the documentation. GitHub repository https://github.com/ClaudioZandonella/trackdown. Full reference: Kothe, E., **Zandonella Callegher, C.**, Gambarota, F., Linkersdörfer, J., & Ling, M. (2021). `trackdown`: Collaborative writing and editing of r markdown (or sweave) documents in google drive [Manual]. https://doi.org/10.5281/zenodo.5167320

`trackdown` overcomes these issues by combining the strengths of literate programming in R with the collaborative features offered by the popular word processor Google Docs.

A.2 Statement of Need

`trackdown` is an R package offering a simple solution for collaborative writing and editing of R Markdown (or Sweave) documents. During the collaborative writing/editing of an `.Rmd` (or `.Rnw`) document, it is important to employ different workflows for computer code and narrative text:

- **Code -** Collaborative code writing is done most efficiently by following a traditional **Git**-based workflow using an online repository (e.g., GitHub or GitLab).
- **Narrative Text** - Collaborative writing of narrative text is done most efficiently using **Google Docs** which provides a familiar and simple online interface that allows multiple users to simultaneously write/edit the same document.

Thus, the workflow's main idea is simple: Upload the `.Rmd` (or `.Rnw`) document to Google Drive to collaboratively write/edit the narrative text in Google Docs; download the document locally to continue working on the code while harnessing the power of Git for version control and collaboration. This iterative process of uploading to and downloading from Google Drive continues until the desired results are obtained (See Figure A.1). The workflow can be summarized as:

> Collaborative **code** writing using **Git** & collaborative writing of **narrative text** using **Google Docs**

Other R packages aiming to improve the user experience during the collaborative editing of R Markdown (or Sweave) documents are available: `redoc` (Ross, 2021) offers a two-way R Markdown-Microsoft Word workflow; `reviewer` (Stringer et al., 2021) allows to evaluate differences between two rmarkdown files and add notes using the Hypothes.is service; `trackmd` (Tyner & Foster, 2021) is an RStudio add-in for tracking changes in Markdown format; `latexdiffr` (Hugh-Jones, 2021) creates a diff of two R Markdown, .Rnw or LaTeX files. However, these packages implement a less efficient writing/editing workflow and all of them, but `latexdiffr`, are no longer under active development. In particular, the `trackdown` workflow has the advantage of being based on Google Docs which offers users a familiar, intuitive, and free web-based interface that allows multiple users to simultaneously write/edit

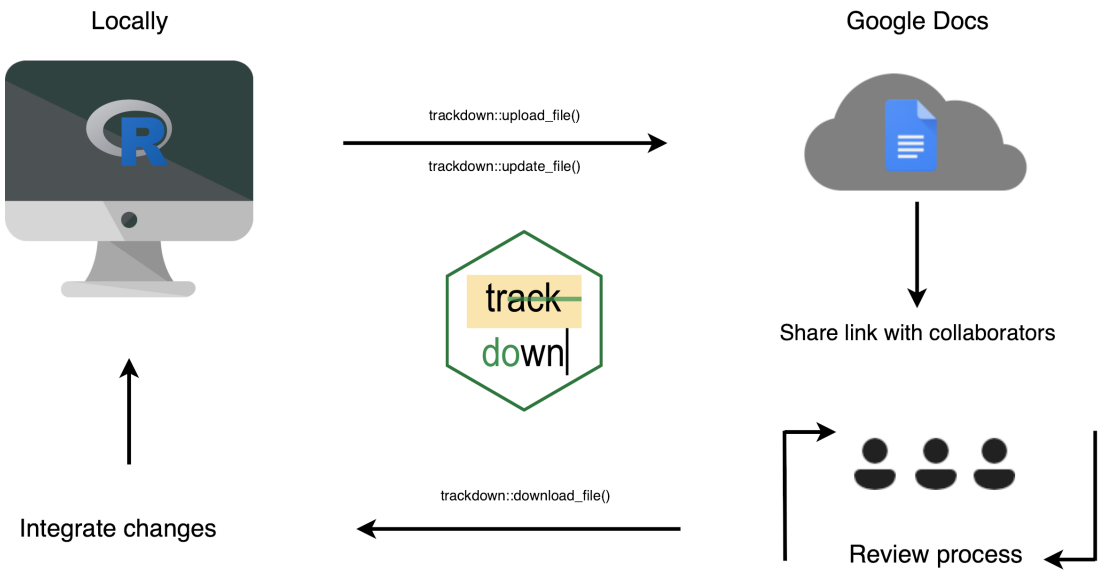

Figure A.1: `trackdown` workflow, collaborative code writing is done locally using Git whereas collaborative writing of the narrative text is done online using Google Docs.

the same document. Moreover, `trackdown` allows anyone to contribute to the writing/editing of the document. No programming experience is required, users can just focus on writing/editing the narrative text in Google Docs.

The package is available on CRAN https://CRAN.R-project.org/package=trackdown and GitHub https://github.com/claudiozandonella/trackdown. All the documentation is available at https://claudiozandonella.github.io/trackdown/.

A.3 Workflow Example

Suppose you want to collaborate with your colleagues on the writing of an R Markdown document, e.g., to prepare a submission to a scientific journal. If you are the most experienced among your colleagues in the usage of R and programming in general, you should take responsibility for managing and organizing the workflow.

A.3.1 Upload File

You create the initial document, for example `My-Report.Rmd`, and upload the file to Google Drive using the function `upload_file()`:

```
library(trackdown)
upload_file(file = "path-to-file/My-Report.Rmd",
            hide_code = TRUE)
```

By executing this command, the `My-Report.Rmd` file is uploaded from your local computer to your Google Drive. Note that `trackdown` adds some simple instructions and reminders on top of the document and, by specifying the argument `hide_code = TRUE` (default is `FALSE`), the header code (YAML) and code chunks are removed from the document displaying instead placeholders of type "`[[document-header]]`" and "`[[chunk-<name>]]`" (See Figure A.2). This allows collaborators to focus on the narrative text.

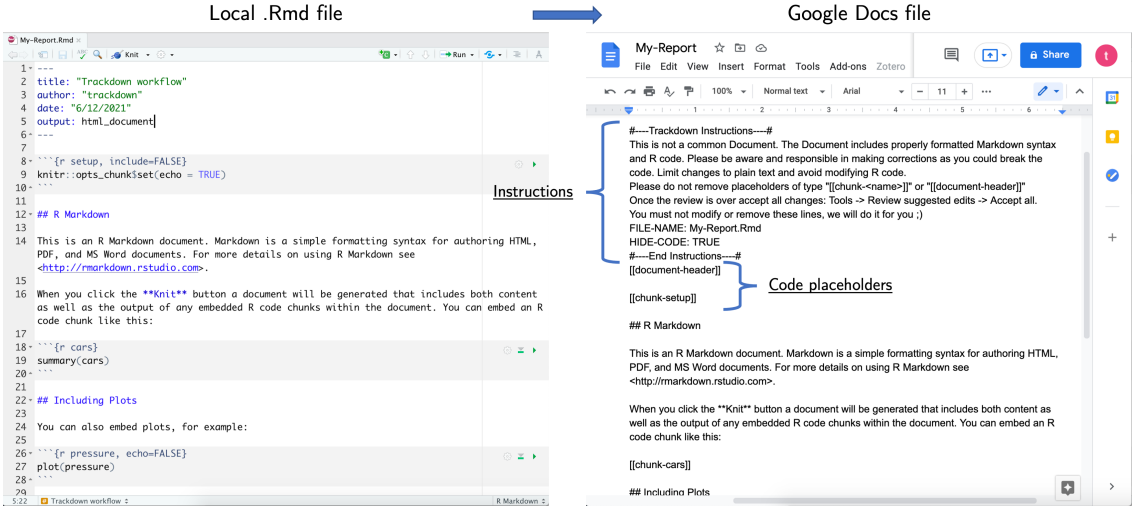

Figure A.2: When uploading a document from your local computer to your Google Drive, `trackdown` adds some simple instructions and reminders on top of the document and, by specifying the argument `hide_code = TRUE` (default is `FALSE`), the header code (YAML) and code chunks are removed and substituted by placeholders.

A.3.2 Collaborate

After uploading your document to Google Drive, you can now share a link to the document with your colleagues and invite them to collaborate on the writing of the narrative text. Google Docs offers a familiar, intuitive, and free web-based interface that allows multiple users to simultaneously write/edit the same document. In Google Docs it is possible to: track changes (incl. accepting/rejecting suggestions); add comments to suggest and discuss changes; check spelling and grammar errors (See Figure A.3).

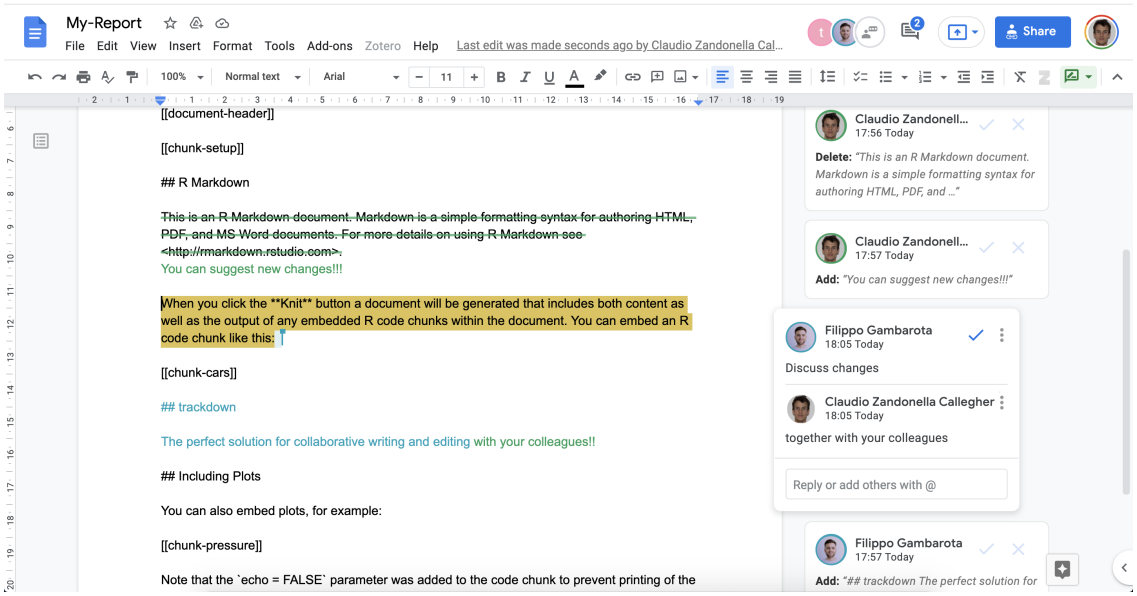

Figure A.3: Example of collaboration in Google Docs using suggestions and comments.

A.3.3 Download File

At some point, you will want to add some code to the document to include figures, tables, and/or analysis results. This should not be done in Google Docs, instead, you should first download the document. Accept/reject all changes made to the document in Google Docs, then download the edited version of the document from Google Drive using the function `download_file()`:

```
download_file(file = "path-to-file/My-Report.Rmd")
```

Note that downloading the file from Google Drive will overwrite the local file.

A.3.4 Update File

Once you added the required code chunks, further editing of the narrative text may be necessary. In this case, you first update the file in Google Drive with your local version of the document using the function `update_file()`:

```
update_file(file = "path-to-file/My-Report.Rmd",
            hide_code = TRUE)
```

By executing this command, the document in Google Drive is updated with your latest local changes. Now you and your colleagues can continue to collaborate on the writing of the document. Note that updating the file in Google Drive will overwrite its current content.

This iterative process of updating the file in Google Drive and downloading it locally continues until the desired results are obtained.

References

Agnoli, F., Melchiorre, F., Zandonella Callegher, C., & Altoè, G. (2021). Stereotype threat effects on Italian girls' mathematics performance: A failure to replicate. *Developmental Psychology*, *57*(6), 940–950. https://doi.org/10.1037/dev0001186

Agresti, A. (2002). *Categorical data analysis*. Wiley.

Ainsworth, M. D., & Bell, S. M. (1970). Attachment, exploration, and separation: Illustrated by the behavior of one-year-olds in a strange situation. *Child Development*, *41*(1), 49–67. https://doi.org/10.2307/1127388

Akaike, H. (1973, November). Information Theory and an Extension of the Maximum Likelihood Principle. In B. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest Akademiai Kiado.

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*. https://doi.org/10.3389/fpsyg.2019.02893

Anderson, S. F. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *The American Journal of Clinical Nutrition*, *110*(2), 280–295. https://doi.org/10.1093/ajcn/nqz058

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Arrison, T. S. (2014). *Responsible science: Ensuring the integrity of the research process*. National Academy of Sciences. United States. https://doi.org/10.2172/1126508

Abstract note: This is the final technical report for DE-SC0005916 Responsible Science: Ensuring the Integrity of the Research Process.

Azzalini, A., Scarpa, B., & Walton, G. (2012). *Data analysis and data mining: An Introduction*. Oxford University Press.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Bartlett, M. (1957). A comment on d. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Bartoń, K. (2019). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018, May 26). *Parsimonious Mixed Models*. arXiv: 1506.04967 [stat]. Retrieved November 22, 2021, from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014, June 23). *Fitting Linear Mixed-Effects Models using lme4*. arXiv: 1406.5823 [stat]. Retrieved March 7, 2019, from http://arxiv.org/abs/1406.5823

Bayarri, M. J., & García-Donato, G. (2007). Extending Conventional Priors for Testing General Hypotheses in Linear Models. *Biometrika*, *94*(1), 135–152.

Berger, J. O., & Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, *91*(433), 109–122. https://doi.org/10.1080/01621459.1996.10476668

Bertoldo, G., Zandonella Callegher, C., & Altoè, G. (in press). Designing Studies and Evaluating Research Results: Type M and Type S Errors for Pearson Correlation Coefficient. *Meta-Psychology*. https://doi.org/10.31234/osf.io/q9f86

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Bowlby, J. (1969). *Attachment and loss*. Basic Books.

Box, G. (1979). Robustness in the Strategy of Scientific Model Building. *Robustness in Statistics* (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brenning, K., Petegem, S. V., Vanhalst, J., & Soenens, B. (2014). The psychometric qualities of a short version of the Experiences in Close Relationships Scale – Revised Child version. *Personality and Individual Differences*, *68*, 118–123.

Bretherton, I. (2010). Fathers in attachment theory and research: A review. *Early Child Development and Care*, *180*(1-2), 9–23. https://doi.org/10.1080/03004430903414661

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., & Munafò, M. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cassidy, J., & Shaver, P. R. (2016). *Handbook of Attachment: Theory, Research, and Clinical Applications* (Third Edition). The Guilford Press.

Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. *JAMA*, *315*(11), 1141–1148. https://doi.org/10.1001/jama.2016.1952

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cook, J., Hislop, J., Adewuyi, T., Harrild, K., Altman, D., Ramsay, C., Fraser, C., Buckley, B., Fayers, P., Harvey, I., Briggs, A., Norrie, J., Fergusson, D., Ford, I., & Vale, L. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess*, *18*(28). https://doi.org/10.3310/hta18280

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

de Santis, F., & Spezzaferri, F. (1999). Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach. *International Statistical Review / Revue Internationale de Statistique*, *67*(3), 267–286. https://doi.org/10.2307/1403706

Dickey, J. M. (1971). The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. https://doi.org/10.1214/aoms/1177693507

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

Du, H., Edwards, M. C., & Zhang, Z. (2019). Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions. *Behavior Research Methods*, *51*(5), 1998–2021. https://doi.org/10.3758/s13428-019-01262-w

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? an fMRI study of social exclusion. *Science*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fisher, R. A. (1938). Presidential address, first Indian statistical congress. *Sankhyā (1933-1960)*, *4*(1), 14–17.

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them, 10. https://doi.org/10.31234/osf.io/hs7wm

Fox, J. (2016). *Applied regression analysis and generalized linear models* (Third Edition). SAGE
OCLC: ocn894301740.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502. https://doi.org/10.1126/science.1255484

Friese, M., & Frankenbach, J. (2020). P-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Fryer, R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240.

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. https://doi.org/10.1177/0146167217729162

Gelman, A. (2019a). Don't calculate post-hoc power using observed estimate of effect size. *Annals of surgery*, *269*(1), e9–e10. https://doi.org/10.1097/SLA.0000000000002908

Gelman, A. (2019b). *From Overconfidence in Research to Over Certainty in Policy Analysis: Can We Escape the Cycle of Hype and Disappointment?* New America. Retrieved May 29, 2020, from http://newamerica.org/public-interest-technology/blog/overconfidence-research-over-certainty-policy-analysis-can-we-escape-cycle-hype-and-disappointment/

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press. https://doi.org/10.1201/b16018

Gelman, A., & Carlin, J. (2013). *Retrospective design analysis using external information* (Unpublished) [Unpublished]. Retrieved April 28, 2020, from http://www.stat.columbia.edu/~gelman/research/unpublished/retropower5.pdf

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American scientist*, *102*(6), 460–466. https://doi.org/10.1511/2014.111.460

Gelman, A., Skardhamar, T., & Aaltonen, M. (2017). Type M Error Might Explain Weisburd's Paradox. *Journal of Quantitative Criminology*. https://doi.org/10.1007/s10940-017-9374-5

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390. https://doi.org/10.1007/s001800000040

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate normal and t distributions*. manual. https://CRAN.R-project.org/package=mvtnorm

Ghosal, R. (2022). Bayesian inference for generalized linear model with linear inequality constraints. *Computational Statistics and Data Analysis*, 17.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n21

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440. https://doi.org/10.1177/0149206314547522

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–1191. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (1999). The Extended Version of the Strengths and Difficulties Questionnaire as a Guide to Child Psychiatric Caseness and Consequent Burden (1999/07/01). *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(5), 791–799. https://doi.org/10.1111/1469-7610.00494

Goodman, S., & Berlin, J. (1994). The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. *Annals of internal medicine*, *121*(3), 200–206. https://doi.org/10.7326/0003-4819-121-3-199408010-00008

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

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511–527. https://doi.org/10.1037/met0000017

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. .-. C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2020, October 20). *A Review of Applications of the Bayes Factor in Psychological Research* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/cu43g

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87. https://doi.org/10.1016/j.jmp.2019.03.004

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Chapman and Hall/CRC. https://doi.org/10.1201/b11158

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Hugh-Jones, D. (2021). *Latexdiffr: Diff 'rmarkdown' files using the 'latexdiff' utility*. manual. https://CRAN.R-project.org/package=latexdiffr

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Pereira, T. V., & Horwitz, R. I. (2013). Emergence of Large Treatment Effects From Small Trials—Reply. *JAMA*, *309*(8), 768–769. https://doi.org/10.1001/jama.2012.208831

Jeffreys, H. (1961). *Theory of probability* (3rd Edition). Oxford University Press.

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

Kato, B. S., & Hoijtink, H. (2006). A Bayesian approach to inequality constrained linear mixed models: Estimation and model selection. *Statistical Modelling*, *6*(3), 231–249. https://doi.org/10.1191/1471082X06st119oa

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. *Journal of Experimental Psychology: General*, *143*(2), 486–491. https://doi.org/10.1037/a0034462

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kothe, E., Callegher, C. Z., Gambarota, F., Linkersdörfer, J., & Ling, M. (2021). *Trackdown: Collaborative writing and editing of r markdown (or sweave) documents in google drive*. manual. https://doi.org/10.5281/zenodo.5167320

Kruschke, J. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2018a). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, *33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/jbh4w

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Laurin, K., Kay, A. C., & Landau, M. J. (2018). Structure and goal pursuit: Individual and cultural differences. *Advances in Methods and Practices in Psychological Science*, *1*(4), 491–494. https://doi.org/10.1177/2515245918797130

Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, *44*(1-2), 187–192. https://doi.org/10.1093/biomet/44.1-2.187

Lindley, D. V. (1972). *Bayesian statistics, a review*. PA: SIAM.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lu, J., Qiu, Y., & Deng, A. (2018). A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12132

Lu, J., Qiu, Y., & Deng, A. (2019). A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 1–17. https://doi.org/10.1111/bmsp.12132

Luce, R. D. (1988). The tools-to-theory hypothesis: Review of G. Gigerenzer and D. J. Murray, "Cognition as intuitive statistics." *Contemporary Psychology*, *33*(7), 582–583.

Marci, T., Moscardino, U., & Altoè, G. (2019). The brief Experiences in Close Relationships Scale - Revised Child version (ECR-RC): Factor structure and invariance across middle childhood and early adolescence. *International Journal of Behavioral Development*, *43*(5), 409–423. https://doi.org/10.1177/0165025418785975

Maslow, A. H. (1966). *The psychology of science: A reconnaissance*. Unknown Publisher.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781107286184

McElreath, R. (**typedirector**). (2020a, September 26). *Science as Amateur Software Development*. Retrieved September 24, 2021, from https://www.youtube.com/watch?v=zwRdO9_GGhY

McElreath, R. (2020b). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor and Francis, CRC Press.

McElreath, R. (2021a, June 15). *Regression, Fire, and Dangerous Things (1/3)*. Elements of Evolutionary Anthropology. Retrieved September 23, 2021, from https://elevanth.org/blog/2021/06/15/regression-fire-and-dangerous-things-1-3/

McElreath, R. (**typedirector**). (2021b, September 10). *Science Before Statistics: Causal Inference*. Retrieved September 28, 2021, from https://www.youtube.com/watch?v=KNPYUVmY3NM

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Moore, D. S. (1998). Statistics among the liberal arts. *Journal of the American Statistical Association*, *93*(444), 1253–1259. https://doi.org/10.1080/01621459.1998.10473786

Mulder, J., & Olsson-Collentine, A. (2019). Simple Bayesian testing of scientific expectations in linear regression models. *Behavior Research Methods*, *51*(3), 1117–1130. https://doi.org/10.3758/s13428-018-01196-9

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)Equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115. https://doi.org/10.1016/j.jmp.2014.09.004

Mulder, J., & Gelissen, J. P. T. M. (2019, April 3). *Bayes factor testing of equality and order constraints on measures of association in social research*. arXiv: 1807.05819 [stat]. Retrieved May 11, 2021, from http://arxiv.org/abs/1807.05819

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1–5. https://doi.org/10.1016/j.jmp.2016.01.002

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., Akrami, N., Ekehammar, B., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 99–138.

O'Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, *73*, 69–81. https://doi.org/10.1080/00031305.2018.1518265

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pastore, M., Lionetti, F., Calcagnì, A., & Altoè, G. (2019). La Potenza è nulla senza controllo - Power is nothing without control. *Giornale italiano di psicologia*, *46*(1-2), 359–378. https://doi.org/10.1421/93796

Pearson, J., & Neyman, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika. A*, *20A*(1/2), 175–240. https://doi.org/10.2307/2331945

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? a tutorial for teaching data testing. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00223

Phillips, B. M., Hunt, J. W., Anderson, B. S., Puckett, H. M., Fairey, R., Wilson, C. J., & Tjeerdema, R. (2001). Statistical significance of sediment toxicity test results: Threshold values derived by the detectable significance approach. *Environmental Toxicology and Chemistry*, *20*(2), 371–373. https://doi.org/10.1002/etc.5620200218

R Core Team. (2021). *R: A language and environment for statistical computing*. manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Ross, N. (2021). *Redoc: Reversible reproducible documents*. manual. https://github.com/noamross/redoc

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021, March 18). *Workflow Techniques for the Robust Use of Bayes Factors*. arXiv: 2103.08744 [stat]. Retrieved August 18, 2021, from http://arxiv.org/abs/2103.08744

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

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

Schooler, J. (2014). Turning the Lens of Science on Itself: Verbal Overshadowing, Replication, and Metascience. *Perspectives on Psychological Science*, *9*(5), 579–584. https://doi.org/10.1177/1745691614547878

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. *Annual Review of Psychology*, *67*(1), 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Stan Development Team. (2020). RStan: The R interface to Stan. http://mc-stan.org/

Stangor, C., & Lemay, E. P. (2016). Introduction to the Special Issue on Methodological Rigor and Replicability. *Journal of Experimental Social Psychology*, *66*, 1–3. https://doi.org/10.1016/j.jesp.2016.02.006

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

Steele, C. M., Spencer, S. J., & Aronson, J. (2002, January 1). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology* (pp. 379–440). Academic Press. https://doi.org/10.1016/S0065-2601(02)80009-0

Stringer, A., Raymond, B., Dulhunty, M., & de Jong, L. (2021). *Reviewer: Improving the track changes and reviewing experience in r markdown*. manual. https://github.com/ropensci-archive/reviewer

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, *11*, 390. https://doi.org/10.3389/fnhum.2017.00390

Timm, A. (2019). *Retrodesign: Tools for type s (sign) and type m (magnitude) errors*. manual. https://github.com/andytimm/retrodesign

Tyner, S., & Foster, Z. (2021). *Trackmd: RStudio addin for tracking document changes*. manual. https://github.com/ropensci-archive/trackmd

van de Schoot, R. (2019). Private communication.

van de Schoot, R., Mulder, J., Hoijtink, H., Van Aken, M. A. G., Semon Dubas, J., Orobio de Castro, B., Meeus, W., & Romeijn, J.-W. (2011). An introduction to Bayesian model selection for evaluating informative hypotheses. *European Journal of Developmental Psychology*, *8*(6), 713–729. https://doi.org/10.1080/17405629.2011.621799

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, Z. Wang, J. Townsend, & A. Eidels (Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Varadhan, R. (2020). *condMVNorm: Conditional multivariate normal distribution*. manual. https://CRAN.R-project.org/package=condMVNorm

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x

Vul, E., & Pashler, H. (2017). Suspiciously high correlations in brain imaging research. *Psychological science under scrutiny* (pp. 196–220). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119095910.ch11

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*, 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102. https://doi.org/10.1016/j.csda.2010.03.016

Wicherts, J. M. (2005). Stereotype Threat Research and the Assumptions Underlying Analysis of Covariance. *American Psychologist*, *60*(3), 267–269. https://doi.org/10.1037/0003-066X.60.3.267

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medicine*, *5*(10), 1–5. https://doi.org/10.1371/journal.pmed.0050201

Zandonella Callegher, C., Toffalini, E., & Altoè, G. (2019). Eliciting effect size - Shiny App (version V1.0.0). https://doi.org/10.5281/zenodo.2564852

Zandonella Callegher, C., Bertoldo, G., Toffalini, E., Vesely, A., Andreella, A., Pastore, M., & Altoè, G. (2021). PRDA: An R package for Prospective and Retrospective Design Analysis. *Journal of Open Source Software*, *6*(58), 2810. https://doi.org/10.21105/joss.02810

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. M. H. , D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

Zondervan-Zwijnenburg, M., van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in psychology*, *8*, 90. https://doi.org/10.3389/fpsyg.2017.00090