UNIVERSITÀ DEGLI STUDI DI PADOVA

**Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata
Dipartimento di Psicologia dello Sviluppo e della Socializzazione**

**Corso di Laurea Magistrale in
Psicologia Sociale, del Lavoro e della Comunicazione**

Tesi di Laurea Magistrale

**Dealing with the replication crisis in psychological science:
The contribution of Type M and Type S errors**

**Affrontare la crisi di replicabilità nelle scienze psicologiche:
Il contributo degli errori di Tipo M e di Tipo S**

Relatore

Prof. Gianmarco Altoè

Correlatore

Prof. Livio Finos

Laureanda: Giulia Bertoldo

Matricola: 1152883

Anno Accademico 2018/2019

Alla mia famiglia

Ringraziamenti

Grazie al prof. Altoè per il sostegno e la supervisione nella realizzazione di questo progetto, per la grande disponibilità e per gli insegnamenti che mi ha trasmesso con passione.

Un ringraziamento speciale va a tutto il gruppo di Psicostat. Lavorare in un ambiente così stimolante, ricco di idee e aperto alla discussione, in un clima di allegria e scambio reciproco è stata un'esperienza unica durante il mio percorso universitario.

Grazie a mamma e papà. Con il vostro esempio di ogni giorno mi mostrate cosa vuol dire affrontare la vita con grinta e spirito positivo. Nella vostra semplicità, mi avete insegnato a non aver paura di far fatica e ad affrontare ogni sfida con serenità. Sono arrivata fino a qui grazie al vostro supporto ed alla fiducia che mi avete dato in qualsiasi progetto abbia deciso di intraprendere. Non vi sarò mai grata abbastanza per tutto questo.

Allo zio Aldo. Grazie per essere sempre presente, pronto in ogni occasione ad aiutarmi a realizzare i miei progetti. Il tuo immancabile sostegno e la tua presenza in famiglia hanno fatto la differenza e mi hanno permesso di arrivare fin qui.

Grazie a Valentina e Mattia per il sostegno e l'interesse che mi dimostrate anche a distanza e per aver portato gioia immensa nella nostra famiglia insieme a Giovanni, Matilde e Sebastian. Valentina, grazie per esserci sempre, sei l'esempio che guardo con ammirazione e che mi ha spronata sin da piccola mettendomi in riga quando avrei voluto fare di testa mia.

Renan, thank you for the immense love and joy you bring in my life. Your invaluable support makes a difference in everything I do.

Thank you to the Junior Researcher Programme for giving me the opportunity to live a life-changing experience. Thank you especially to all the people I worked with during the last year, you are a tremendous source of inspiration for me.

Thank you to all my friends who have been close to me during this journey.

Table of Contents

Riassunto.....	1
Summary.....	3
Chapter 1 Statistical Inference	5
1.1 Statistical Inference.....	5
1.2 Fisher.....	7
1.3 Neyman and Pearson.....	9
1.4 The Hybrid: NHST.....	13
1.5 Statistical Power.....	16
1.6 Effect Sizes	17
Chapter 2 The Replication Crisis in Psychological Science	21
2.1 The Crisis of Confidence in Published Research Findings.....	21
2.2 Replicability	23
2.2.1 Replicability and Reproducibility: Definitions	23
2.2.2 Difficulties When Interpreting Replication Studies' Results.....	24
2.3 An Example of Replication Project in Psychology.....	26
2.4 Questionable Research Practices (QRPs)	27
2.5 Aren't Scientists Interested in "Finding Things Out"?	32
2.6 Changing Mindset, Practices and Incentives	33
Chapter 3 Errors of Magnitude and Sign: Type M and Type S.....	37
3.1 The Statistical Significance Filter in Noisy Studies.....	37

3.2 What is “Noise”	39
3.3 The Winner’s Curse	42
3.4 Type M and Type S Errors.....	44
3.4.1 Definitions.....	44
3.4.2 Plausible Effect Size	44
3.4.3 Type M, Type S and Statistical Power.....	45
3.5 <i>Design Analysis</i> : Prospective and Retrospective	48
3.5.1 Case Study.....	49
3.5.2 <i>Prospective Design Analysis</i>	50
3.5.3 <i>Retrospective Design Analysis</i>	52
Chapter 4 Type M Error Might Explain the Decline Effect	55
4.1 The Decline Effect: Definition and Types	55
4.2 Case Study.....	59
4.2.1 The Original Study.....	59
4.2.2 Study 4 and the Registered Replication Report 1	61
4.2.3 Study 1 and the Replication Registered Report 2	65
Chapter 5 Conclusions.....	69
References	73
Appendices.....	81

Riassunto

Una parte considerevole della ricerca in psicologia utilizza metodi quantitativi per la progettazione e l'analisi degli studi. La statistica, in particolar modo quella inferenziale, ricopre un ruolo centrale nel processo di ricerca, permettendo di passare dalla teoria psicologica alla quantificazione e valutazione delle ipotesi di interesse. Nell'ultimo decennio diversi studi si sono posti l'obiettivo di verificare la possibilità di ritrovare gli stessi risultati di studi presenti in letteratura replicandone la procedura. Le conclusioni furono poco incoraggianti, difatti molti effetti non risultarono essere così forti come negli studi originari ed in alcuni casi gli studi di replica non ritrovarono alcun effetto. Questi risultati hanno portato la comunità scientifica a riconoscere una crisi di replicabilità in psicologia. Tuttavia, visto il ruolo centrale della statistica nel processo di ricerca e l'osservazione di un simile andamento anche in altre scienze, si è cominciato a parlare di “*statistical crisis in science*” (Gelman & Loken, 2014).

Il presente lavoro ha l'obiettivo di analizzare la crisi di replicabilità in psicologia con un particolare focus sui processi di inferenza statistica. L'obiettivo principale è mettere in evidenza i rischi da tenere in considerazione quando si effettuano test inferenziali per la valutazione delle ipotesi in ottica Frequentista. Oltre ai classici errori di Tipo I e di Tipo II, altri due errori non comunemente considerati sono l'errore di Tipo M (*magnitude*) e l'errore di Tipo S (*sign*), riguardanti la grandezza e la direzione degli effetti.

Il primo capitolo introduce il *Null Hypothesis Significance Testing* (NHST), ovvero l'approccio di inferenza statistica più diffuso nelle scienze sociali. A tal proposito viene utilizzata una prospettiva storica e vengono presentati anche gli approcci degli statistici Ronald Fisher, Jerzy Neyman ed Egon Pearson. Il secondo capitolo presenta la crisi di replicabilità in psicologia analizzandone le origini, i fattori che hanno contribuito alla crisi e le soluzioni proposte per un cambio di direzione. Il terzo capitolo analizza il ruolo degli errori di Tipo M e di Tipo S nella crisi di replicabilità. Studi con alta probabilità di commettere questi due tipi di errori potrebbero fornire

stime di effetti la cui grandezza è esagerata e/o nella direzione sbagliata. A questo scopo, vengono presentate due tipi di analisi con l'obiettivo di esaminare queste tipologie di errori prima di condurre uno studio (*prospective design analysis*) oppure una volta che lo studio è già stato condotto (*retrospective design analysis*). Il quarto capitolo si propone di collegare le considerazioni sugli errori di Tipo M e di Tipo S al *decline effect*, ovvero l'osservazione che la grandezza degli effetti diminuisce con ripetuti studi di replica. Nonostante ci possano essere molteplici spiegazioni al *decline effect*, una di queste è che lo studio originario riporti un effetto sovrastimato. Si propone un *case study* per mostrare come una *retrospective design analysis* può fornire informazioni sulla probabilità di errore di Tipo M e di Tipo S nello studio originale e fornire supporto all'ipotesi di sovrastima. Il quinto capitolo riassume i contributi degli errori di Tipo M e di Tipo S nella crisi di replicabilità ed il ruolo della *design analysis* nella progettazione di studi ed analisi dei risultati.

Summary

A considerable part of research in psychology employs quantitative methods to design and analyze studies. Statistics, and in particular inferential statistics, plays a central role in the research process, making it possible to move from psychological theory to the quantification and evaluation of hypotheses. Over the last decade, researchers conducted replications of some studies published in the literature. The main aim was to check whether the same results could be found by repeating the same procedure. Replication studies' results were not very encouraging. In fact, many effects were not as strong as in the original studies and, in some cases, replication studies did not find any effect. These results led the scientific community to recognize a replicability crisis in psychology. However, given the central role of statistics in the research process and the observation of a similar trend in other sciences, the scientific community started to talk about a "statistical crisis in science" (Gelman & Loken, 2014).

The present work aims to analyze the replicability crisis in psychology with a focus on statistical inference. The main objective is to highlight the risks to beware when performing hypotheses tests in a Frequentist framework. In addition to the classic Type I and Type II errors, two other errors that are not commonly considered are Type M error (magnitude) and Type S error (sign), concerning the size and direction of the effects.

The first chapter introduces Null Hypothesis Significance Testing (NHST), the prevalent approach to statistical inference in the social sciences, following a historical perspective and presenting also the approaches of the statisticians Ronald Fisher, Jerzy Neyman and Egon Pearson. The second chapter discusses the replicability crisis in psychology with an analysis of the origins, the factors that contributed to the crisis and solutions proposed for a change of direction. The third chapter analyzes the role of Type M and Type S errors in the replicability crisis. Studies with a high probability of committing these two types of errors could provide estimates of effects that are exaggerated and/or in the wrong direction. Two types of analysis are presented to examine these

Summary

errors before conducting a study (*prospective design analysis*) or once the study has already been conducted (*retrospective design analysis*). The fourth chapter aims to link Type M and Type S errors with the *decline effect*, which is the observation that the magnitude of effects decreases with repeated replications. Although there may be multiple reasons behind the decline effect, a possible explanation is that the original study overestimated the effect. A case study illustrates how a *retrospective design analysis* of the original study can provide information on the probability of Type M and Type S errors and give support to the hypothesis of overestimation. The fifth chapter summarizes the contributions of Type M and Type S errors to the replication crisis and the role of a *design analysis* in planning studies and analyzing results.

Chapter 1

Statistical Inference

*"If statisticians agree on one thing,
it is that scientific inference
should not be made mechanically."
(Gigerenzer & Marewski, 2015)*

In this chapter, we describe three approaches to statistical inference: the one of Fisher, the one of Neyman and Pearson and Null Hypothesis Significance Testing (NHST). It has been proposed that NHST is an incoherent framework which was born in statistical textbooks for social scientists as a mixture of Fisher's Null Hypothesis Testing and Neyman and Pearson Decision Theory (Gigerenzer, 1993). Even though the similarities between the last two frameworks, they were developed for different purposes and provided different interpretations of concepts with the same name (e.g., statistical significance). Thus, combination in NHST produced some incoherencies. When NHST is presented, it is not usually mentioned how it originated and it appears to be the only method to statistical inference. This could partially explain some statistical misconceptions and the neglect of notions such as statistical power and effect sizes, which will be the focus of discussion at the end of the chapter.

1.1 Statistical Inference

Researchers in Psychology who choose quantitative methods to tackle research questions make abundant use of statistical tools. The discipline of statistics offers a toolbox to analyze data which includes, for instance, descriptive statistics, exploratory data analysis, Fisher's null hypothesis testing, Neyman-Pearson hypotheses testing, Wald's sequential analysis and Bayesian statistics (Gigerenzer, Krauss, & Vitouch, 2004). Despite this wealth of methodologies, it is

common among consumers of statistics in the social sciences to know only one side of the toolbox, known as Null Hypothesis Significance Testing (NHST). NHST is an incoherent combination of Fisher's and Neyman-Pearson frameworks (Gigerenzer, 2004). Both these approaches are part of inferential statistics, which allows inferring properties of populations by analyzing samples taken from those populations.

However, this focus on statistical inference has not always been the rule in Psychology. As noted by Gigerenzer (1993) "Piaget worked out his logical theory of cognitive development, Kohler the Gestalt laws of perception, Pavlov the principles of classical conditioning, Skinner those of operant conditioning, and Bartlett the theory of remembering and schemata – all without rejecting null hypotheses" (p. 311). Changes started to happen around the years 1940 and 1955 with the inference revolution in the United States (Gigerenzer & Marewski, 2015). During those years, textbooks for social scientists typically portrayed NHST as "the method" of scientific inference, and journal policies started to evaluate research quality based on the level of significance (Gigerenzer, 1993). This contributed to an understanding of statistical inference as a mechanic sequence of steps or ritual, become known as Null Hypotheses Significance Testing. (Gigerenzer, 2004).

This mechanical ritual aims to resolve a conflict between incompatible approaches. On one hand, there is the desire to know the probability of a hypothesis given the data $P(H/D)$, which is possible by adopting Bayesian statistics. On the other hand, there are the rules of Neyman-Pearson hypotheses testing that allow to make a decision among competing courses of actions (accept one hypothesis or the other) based on the p -value, which is the probability of the data obtained, or more extreme data, under the null hypothesis or $P(D/H_0)$. Researchers who need to publish resolve this conflict by stating only one hypothesis, computing the significance level after the experiment using a conventional Type I error rate (α) and ignoring the Type II error rate (β) and sample size planning. One issue with this procedure is that it invites no judgment and Fisher, Neyman and Pearson would have all considered this approach as bad statistics (Gigerenzer, 2004).

Ignorance about the statistical toolbox and the adoption of a mechanical ritual for inferences has contributed to the misunderstanding of concepts such as p -values (Haller & Krauss, 2002; Oakes, 1986), and the neglect of other important concepts such as power, effect sizes and confidence intervals (Sedlmeier & Gigerenzer, 1989). The next paragraphs will present Fisher's Null Hypothesis Testing, Neyman-Pearson Decision Theory and Null Hypothesis Significance Testing (NHST) and will conclude with a discussion on statistical power and effect sizes.

1.2 Fisher

Fisher proposed a method of statistical inference called *Null Hypothesis Testing* or *Significance Testing* (Table 1). In this method, the researcher specifies a hypothesis called *Null*, not necessarily because it postulates an effect equal to zero, which is instead called the nil hypothesis. The null hypothesis refers to the hypothesis that the researcher is trying to nullify (Gigerenzer, 2004). The observations collected in a study are compared to the observations expected if the null hypothesis was to be the real state of the world. The aim is to estimate how much the observations deviate from those predicted by the null hypothesis. If the data are far from plausible values under the null hypothesis, then it is possible to reject the null hypothesis as a plausible state of the world.

Table 1

Fisher Significance Testing or Null Hypothesis Testing

-
1. Set up a statistical null hypothesis. The null need not to be a nil hypothesis (zero difference).
 2. Report the exact level of significance (e.g., $p = .011$ or $.051$). Do not use a conventional 5% level (e.g., $p < .05$), and do not talk about accepting or rejecting hypothesis.
 3. Use this procedure only if you know very little about the problem at hand
-

Note. Content adapted from Gigerenzer, Krauss, & Oak (2004).

Fisher introduced the concept of *significance level* and p -values to indicate how much extreme the data should be to reject the null hypothesis. In the book the “Design of experiments” (1936), Fisher defined the level of significance as equal to 5%, or sometimes 1%. If the probability of the data collected, under the null hypothesis, was lower than 5% (i.e., p -value $< .05$), then results

could be considered *statistically significant*, leading to the rejection of the null hypothesis.

However, Fisher offered contrasting interpretations of the significance level at different points during his life (Gigerenzer, 1993). First, Fisher argued that this 5% should be used in all situations. Hence, results below the significance level could be reported as $p < .05$ regardless of the exact p -value. Later, Fisher proposed a different interpretation asserting that it does not make sense to keep the same significance level in different contexts. What scientists should do instead, is to report the exact p -value obtained in a study (Gigerenzer, 1993).

What about non-statistically significant results? Fisher changed his views over time on how these should be interpreted. First, he argued that non-significant results should be ignored, whereas later on, he claimed that non-significant results could be interpreted as *confirming* the null hypothesis, but importantly they could never *establish* a null hypothesis (Gigerenzer, 1993).

Fisher held an epistemic interpretation of probability. In other words, he thought that the level of significance of a result could give information about the degree of belief in the null hypothesis (Gigerenzer, 1993). Hence, the exact level of significance should be interpreted as “a property of the data, that is, a relation between a body of data and a theory” (Gigerenzer, Krauss, & Vitouch, 2004, p. 11). This point is one of the main differences with Neyman and Pearson framework which is instead a decision theory where the level of significance α denotes the relative frequency of making one type of wrong decision, and there is no intention to give an epistemic interpretation of a hypothesis as true or false.

In conclusion, Fisher suggested Significance Testing as adequate only when we know nothing or little about the phenomena and not as a procedure that can be applied in every situation in scientific research (Gigerenzer et al., 2004). This situation is rare in psychological science, where usually there is a wealth of competing theories and published studies.

It is important to remark that within Null Hypothesis Testing only one hypothesis is specified (Figure 1). This is one of the factors that led Jerzy Neyman and Egon Pearson to elaborate their approach to statistical inference, where at least two competing hypotheses are specified and where also the concepts of statistical power and effect sizes are included (Gigerenzer, 2004).

Figure 1. Null Hypothesis Testing. Sampling distribution of the test statistics, for one-tailed test and a p -value of 0.02. In Fisher's approach, the null hypothesis does not need to be a nil hypothesis. Adapted from 'Show me Shiny, gallery of R web apps', 2019, <https://casertamarco.shinyapps.io/power/>

1.3 Neyman and Pearson

Neyman and Pearson framework can be seen as a decision theory that allows answering the question "What should I do now?" (Royall, 2017). In order to answer this question, an objective relative frequency interpretation of probability is used. For instance, the probability of obtaining head, when tossing a coin is the relative frequency of heads in an infinite number of coin tosses. The infinite number of tosses is the reference class or collective and it specifies how the data are obtained. For example, one reference class can be tossing a coin with the right arm, whereas a different reference class is tossing a coin with the right arm moving in a precise way in particular wind conditions (Dienes, 2008). The probability of obtaining head associated with the first reference class is different from the second (e.g., 50% for the first and 80% for the second). It is important to note that in this case the probability refers to a collective and not to a single event. In

other words, it is reasonable to say how many heads will be obtained in the long run with a certain procedure, but it is not possible to define the probability of a single event (e.g., next toss being head; Dienes, 2008).

The reference class plays a role in establishing how to build the sampling distributions of the test statistics (Dienes, 2008). In the Neyman and Pearson framework, the researcher sets two competing hypotheses (H_1 and H_2), one is called the “null hypothesis” and the other the “alternative hypothesis”. Then, the reference class is used to compute the sampling distribution of the test statistics if H_1 were to be the real state of the world and if H_2 were to be the real state of the world (Figure 2). Given that the aim is to take a decision, the researcher has to set a threshold that establishes when a decision is taken over the other (i.e., a region of acceptance of H_1 and a region of acceptance of H_2). This threshold depends on the probability that the research is willing to accept of committing two possible types of errors when repeating this experiment a number of times that tends to infinite:

1. to accept H_2 when H_1 is the real state of the world, called Type I error.
2. to accept H_2 when H_1 is the real state of the world, called Type II error.

The probability of committing a Type I error is indicated with α . For example, if α is set at 5%, it means that the researcher accepts to make a Type I error no more than 5% of the times when repeating the experiment a number of times that tends to infinite. This probability allows establishing the *critical value* above which H_2 is accepted, and under which H_1 is accepted. This area is represented in Figure 2 as the area under H_1 where the decision rule states to “accept H_1 ”.

The probability of committing a Type II error is indicated with β and it depends, among other things, on the size of the effect under study. The area under H_2 where the decision rule states “accept H_1 ” represents the probability of a Type II error. For instance, if the researcher sets $\beta = 20\%$, he accepts to commit a Type II error no more than 20% of the times when repeating the experiment a number of times that tends to infinite.

To sum up, this approach allows to take decisions, but this means that each time a decision is taken, it can either be correct or wrong. What is important then, is to assure that the frequency with which wrong decisions are made over time are maintained under a certain level. Hence, the need to establish the two error rates α and β .

Once α and β are settled, it is possible to know what the conditional probability is of making the correct decision in the long run. This could either be to accept H_1 when this is the real state of the world ($1 - \alpha$), or conversely, to accept H_2 when this is the real state of the world. The latter is called the *statistical power* of the test ($1 - \beta$) and it plays a central role in the current discussion about the replicability crisis in Psychology. Table 2 summarizes these conditional probabilities and the possible decisions in Neyman and Pearson's framework.

Figure 2. Neyman and Pearson's approach. Sampling distributions of the test statistics under H_1 (i.e., Null Hypothesis) – grey line, and H_2 (i.e., Alternative Hypothesis) – blue line. There are two possible decisions: accept H_1 or accept H_2 . The red area represents the Type I error rate α (e.g., 0.05), and the dark grey area represents the Type II error rate (e.g., 0.20). Power is the combination of the light grey area and red area under the blue curve ($1 - \beta$). Adapted from 'Show me Shiny, gallery of R web apps', 2019, <https://casertamarco.shinyapps.io/power/>

Table 2

Possible Decisions in the Neyman-Pearson Framework

		<i>True state of the world</i>	
<i>Decision</i>		<i>H₁ is True</i>	<i>H₂ is True</i>
	<i>Accept H₁</i>	Correct decision (1 – α)	Type II error (β)
	<i>Accept H₂</i>	Type I error (α)	Correct decision Statistical Power (1 – β)

This framework is especially suitable in contexts such as quality control. For example, imagine a factory that produces disks which have to be of a certain diameter. It is not possible to check if every produced disk has the required diameter (e.g., 5 cm) or a bigger one. However, samples can be taken from the daily production to test two hypotheses:

$$H_1: \text{diameter} = 5 \text{ cm}$$

$$H_2: \text{diameter} > 5 \text{ cm}$$

In order to test these hypotheses, we would have to take a sample of disks (e.g., 100 disks) and measure the diameter of each one. Then, we could calculate the “mean diameter” for that daily production (e.g., 5.3 cm). To see if the mean diameter of the sample is greater than 5 cm or if this is not a surprising result under H_1 , we could perform a *t-test* and see if the *t* - statistics lies in the region of acceptance of H_1 or H_2 . If H_1 is accepted, then the production can continue, otherwise, if H_2 is accepted, the production has to be stopped to check the machineries producing the disks.

It is possible to note that in this paradigm, the probability of obtaining a value equal to or greater than the critical value (*p*-value) is used to make a decision based on the sample. It is then not important the precise *p*-value but only if it is greater or smaller than the critical value, and then act as if the hypothesis that is accepted was to be the true state of the world. Thus, this approach does not tell us anything about the truth of a hypothesis but only how to act based on the assumption that one hypothesis is true (Table 3).

Table 3

Neyman and Pearson Hypotheses Testing

-
1. Set up two statistical hypotheses, H_1 and H_2 , and decide about α , β , and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
 2. If the data falls into the rejection region of H_1 , accept H_2 ; otherwise, accept H_1 . Note that accepting a hypothesis does not imply that you believe in it; it only means that you act as if it were true.
 3. The usefulness of the procedure is limited to situations in which you have a disjunction of hypotheses (e.g., $\mu = 8$ or $\mu = 10$ is true) and in which the scientific context can provide the utilities that enter the choice of α and β .
-

Note. Adapted from Gigerenzer, Krauss, & Oak (2004).

Fisher considered this approach inappropriate for scientific problems where usually it is not even possible to repeatedly perform random sampling from a precise population (Gigerenzer, 1993). However, Neyman and Pearson considered Fisher's approach useless in those cases where some tests had less power than the alpha level (Gigerenzer, 1993). Something both parties agreed upon was that inference should not be an automatic process, like following a recipe in a cookbook, but it should instead be based on informed judgments (Gigerenzer, 1993). Unfortunately, this is the general attitude promoted by the hybrid version of the two approaches that became popular among social scientists.

1.4 The Hybrid: NHST

The approaches of Fisher and Neyman and Pearson are different in many respects and one of the most important is their objective (i.e., a decision-theory vs. a way of knowing something about the truth of a hypothesis). When statistical textbooks for social scientists started to be written, these incompatible frameworks were melted together by picking some aspects from both approaches and producing what is known as NHST: Null Hypotheses Significance Testing (Gigerenzer, 1993). This approach was presented without reference to the original authors (i.e., anonymity) and giving the impression of one monolithic method of statistical inference (Gigerenzer, 2004). Moreover, this

method became institutionalized not only in university curricula but also in journal policies to evaluate research quality for publishing (Gigerenzer, 1993).

This approach is incoherent in many ways, but especially because it uses the level of significance of Fisher's approach in a sort of Neyman and Pearson framework where the aim is to decide between two competing hypotheses. We refer here to the Null Hypothesis as H_0 and the Alternative Hypothesis as H_1 . One problem with this approach is that H_1 is not formalized, for example, by specifying an effect size. As a consequence, the probability of a Type II error and the associated statistical power can be easily ignored. Instead, what is usually performed is a mechanic procedure which Gigerenzer et al. (2004) called the "Null Ritual" and which steps are reported in Table 4.

Table 4

NHST (Null Hypothesis Significance Testing)

-
1. Set up a statistical null hypothesis of "no mean difference" or "zero correlation". Don't specify the predictions of your research hypothesis or of any alternative hypotheses.
 2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis.
 3. Always perform this procedure.
-

Note. Adapted from Gigerenzer, Krauss, & Oak (2004).

Furthermore, the language changes from the Neyman-Pearson approach where it was possible to accept one hypothesis or the other. In NHST it is never possible to accept a hypothesis, but only to *reject the null hypothesis* or *fail to reject the null hypothesis* (Table 5; Figure 3). These decisions are made by looking at the p -value obtained from the test statistic with respect to the critical value. If the p -value is greater than the critical value, then H_0 cannot be rejected and the judgment should be suspended. If the p -value is lower than the critical value, then H_0 can be rejected. However, two common misinterpretations consist in going beyond the "reject H_0 " and "fail to reject H_0 " conclusions by claiming from a statistically significant result (i.e., p -value lower than the critical value) that the alternative hypothesis is true or to claim from a non-statistically

significant result (p -value greater than the critical value) that the null hypothesis is true (e.g., “there was no difference between the two groups” or “there was no association”).

Figure 3. Null Hypotheses Significance Testing approach (NHST). Sampling distribution of the test statistics under H_0 (i.e., Null Hypothesis) with two possible decisions: reject H_0 or fail to reject H_0 . This is a two-tailed test and α is set at 0.05. The significance level is used to establish the critical values. If the p -value is greater than the critical values, then H_0 is rejected. In NHST, the null hypothesis is usually set as a nil hypothesis. Adapted from ‘Show me Shiny, gallery of R web apps’, 2019, <https://casertamarco.shinyapps.io/power/>

Table 5

Possible decisions in the NHST framework

		True state of the world	
		H_0 is True	H_1 is True
Decision	Fail to reject H_0	Correct decision ($1 - \alpha$)	Type II error (β)
	Reject H_0	Type I error (α)	Correct decision Statistical Power ($1 - \beta$)

1.5 Statistical Power

In Null Hypothesis Significance Testing the concept of statistical power can be easily neglected because only the null hypothesis is formalized, whereas the alternative hypothesis is not specified in statistical terms. The problem is that the null hypothesis can become a “straw man” where its rejection is taken as a way to accept the alternative hypothesis (Gelman & Carlin, 2017). In reality, if an alternative hypothesis is introduced and tested, then the Type II error rate should be controlled, and statistical power should be considered.

Not reflecting on power could lead to perform underpowered studies without the awareness of the associated risks. One clear consequence of conducting an underpowered study is that the probability of being able to reject H_0 is very low, if H_0 is actually false. This means that experiments could be “dead on arrival” (Gelman, 2016) in the sense that even if the effect of interest exists, the study will have a low probability to find it. A related misinterpretation is that if a statistically significant result is found despite the low probability of this to happen (i.e., low power), then researchers might consider the results even more remarkable. This error is related to the “what does not kill statistical significance makes it stronger fallacy” (Loken & Gelman, 2017) according to which sampling error, measurement error, extraneous variables and researchers’ degrees of freedom, can easily inflate the probability of finding a p -value lower than the critical value. Hence, results from noisy studies, such as underpowered studies, should be carefully scrutinized (Gelman & Loken, 2014; Loken & Gelman, 2017; Simmons, Nelson, & Simonsohn, 2011).

These misunderstandings contributed to a plague of “significant findings” in the literature, that however originated from underpowered studies (Sedlmeier & Gigerenzer, 1989). In 1962, Jacob Cohen, a statistician and psychologists, published the first systematic review of the power of studies published in the volumes of the Journal of Abnormal and Social Psychology in 1960. He found that studies had 18%, 48% and 83% power to detect small, medium and large effect sizes, respectively (Sedlmeier & Gigerenzer, 1989). Sedlmeier & Gigerenzer (1989) examined the 1984 volume of the same journal and found that the median power was 16%, 46%, and 89% for detecting

small, medium and large effect sizes respectively. The authors then concluded that: “As a general result the power has not increased after 24 years” (Sedlmeier & Gigerenzer, 1989, p.309). More recent studies do not indicate that the situation has improved. Button et al. (2013) found the average power to be around 20% in a sample of studies in neuroscience. Smaldino and McElreath (2016), collected results from reviews on statistical power in the social, behavioral and biological sciences published between 1960 and 2011 and found that the mean power for the detecting small effect sizes (i.e., Cohen’s d of 0.20) was about 20%. Furthermore, across the years, there was no increase in statistical power (Figure 4).

Figure 4. Results of reviews on the statistical power of studies published between 1960 and 2011. The mean power is 0.24 and the black line indicates that there has not been an improvement in the years. Reprinted from ‘The Natural Selection of Bad Science’, by P.E. Smaldino and R. McElreath, 2016, *Royal Society open science*, 3(9), p.6.

1.6 Effect Sizes

The concept of power becomes very natural when thinking that what researchers are trying to do with a study is to measure quantities. This is very different from being merely interested in testing if an effect is either present or absent, which is instead the usual conclusion drawn from an exclusive look at p -values. Effects in Psychology can be, for example, levels of symptoms of a pathology, Intelligence Quotient (IQ) levels, levels of empathy, math abilities. These are latent

constructs, which means that they cannot be readily measured like physical quantities such as temperature, weight, and height. In fact, physical quantities are expressed in comparison to a unit of measurement that is meaningful and readily interpretable, like degrees Celsius (or Kelvin, or Fahrenheit), kilograms and meters.

Latent constructs are measured using effect sizes that are standardized, in the sense that they are expressed in comparison with a measure of variability (i.e., standard deviation). For example, a study measuring the *IQ* level of a group of students with a validated test, expresses the raw scores from the questionnaire (i.e., unstandardized measure) into standardized scores by subtracting from the mean of the group (IQ_{sample}), the mean *IQ* of the population to which the sample belongs ($IQ_{population}$), divided by the standard deviation of the population ($SD_{population}$).

$$IQ_{sample} = \frac{IQ_{sample} - IQ_{population}}{SD_{population}}$$

However, when performing experiments, researchers are mostly interested in the relationship between variables or differences between groups. Two families of effect sizes have been proposed to deal with these problems: the *d* family, to express group differences and the *r* family, to express associations between variables. One common effect size measure in the *r* family is the correlation coefficient. Whereas a common effect size in the *d* family is Cohen's *d* (Ellis, 2010):

$$Cohen's\ d = \frac{M_1 - M_2}{SD_{pooled}}$$

Where M_1 is the mean of one group, M_2 is the mean of the other group and SD_{pooled} is the pooled standard deviation:

$$SD_{pooled} = \sqrt{\frac{\sum_i (X_{i,1} - M_1)^2 + \sum_i (X_{i,2} - M_2)^2}{n_1 + n_2 - 2}}$$

Where $X_{i,1}$ is the i -th observation in the first group, $X_{i,2}$ is the i -th observation in the second group, n_1 is the sample size of the first group and n_2 is the sample size of the second group.

Cohen (1988) provided some benchmarks to interpret effect sizes suggesting indicative values of d for “small”, “medium”, and “large” effect sizes. A *small effect size* corresponds to $d = .2$ and it refers to small differences that are difficult to detect. For example, approximately the difference in mean height between 15- and 16-year-old girls. A *medium effect size* corresponds to $d = .5$ and it refers to differences that are “large enough to be visible to the naked eye” (Cohen, 1988, p.26). For example, the magnitude of the difference in height between 14- and 18-year-old girls. Finally, a *large effect size* corresponds to $d = .8$ and it refers to very obvious differences such as the mean difference in height between 13- and 18- year-old girls. Even though these benchmarks might be useful to understand how to interpret Cohen’s d , they should not be interpreted equally in different contexts. For example, “small effect sizes can have large consequences, such as an intervention that leads to a reliable reduction in suicide rates with an effect size of $d = 0.1$ ” (Lakens, 2013, p. 3).

Effect sizes are important because they allow moving the focus from accepting or rejecting hypotheses to accurately estimate the parameters of interest. Test statistics used to compute p -values can be converted into standardized effect size measures and the uncertainty can be expressed using confidence intervals. For example, a 95% confidence interval indicates that, if an experiment were to be repeated a number of times that tends to infinite, 95% of the obtained confidence intervals would contain the true population value (Greenland et al., 2016).

There are two important points regarding the relationship between statistical significance and effect sizes. The first one is that it is possible to obtain small p -values associated with effects that are too small to be of practical importance. The second one is that it is also possible to obtain small p -values associated with large effect-sizes that however originates from underpowered

studies. In the second situation, the estimated effect sizes are too big to be true, the confidence intervals are large and small p -values can be obtained only because the true effect sizes are overestimated (i.e., Type M error).

In conclusion, considerations about effect sizes and power of statistical tests are of fundamental importance to correctly interpret studies' results. Furthermore, two other types of errors can be derived when the focus is on estimating effect sizes and not only on rejecting hypotheses. These are called Type M (magnitude) and Type S (sign) errors (Gelman & Carlin, 2014) and will be the focus of the following chapters (Chapter 3 and Chapter 4). The crisis of unreplicable results in psychological science had the benefit of promoting increased awareness about these issues among consumer of statistics and social scientists (Chapter 2).

Chapter 2

The Replication Crisis in Psychological Science

*“If wrong ideas or result are not challenged,
we let down future generations who will try
to build on a research base that is not a solid foundation.
Worse still, when the research findings have practical
applications in clinical or policy areas,
we may allow wrongheaded interventions or policies
to damage the well-being of individuals or society.”
(D.V.M. Bishop, 2018)*

In this chapter, we will present the origins of the crisis of confidence in the published literature in psychological science. We will then discuss the new emphasis on replicability and reproducibility of research findings with the discussion of a replication project (Open Science Collaboration, 2015). Then, we will explore various factors hypothesized to be at the base of the low replication rate (e.g., Questionable Research Practices). The chapter will conclude describing some new practices that have been proposed for a change of direction, such as studies' preregistration and platforms to share data and materials.

2.1 The Crisis of Confidence in Published Research Findings

In 2005 a paper was published in PLoS Medicine with the title “Why Most Published Research Findings Are False” (Ioannidis, 2005). This could be considered the beginning of the increased awareness of problems with the published literature and the start of the so-called “crisis of confidence” in some scientific fields such as medicine and psychology. The author posits the foundation of his strong statement in the “ill-founded strategy of claiming conclusive research findings solely based on a single study assessed by formal statistical significance, typically for a p -value less than 0.05” (Ioannidis, 2005, p. 696). Although it is not possible to establish with certainty which finding is true and which is false, it is possible to consider factors that influence the post-study probability that a research finding is true, also called Positive Predictive Value (PPV).

The PPV increases with increasing power, decreasing bias and decreasing testing by several independent teams. Bias is the “combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced” (Ioannidis, 2005, p. 697). It includes performing multiple analyses to find desired outcomes or reporting partial data, analyses or conditions of a study. All these behaviors go under the term “Questionable Research Practices” (QRPs). Testing by independent teams refers to the widespread attitude of many laboratories to tackle similar questions separately without considering the evidence altogether with a preference to focus on isolated findings. Table 6 presents some corollaries that can be derived from the interplay of these two factors (Ioannidis, 2005). The corollaries can be used to gain insights into the state of a research field from the characteristics of the studies conducted in it.

Table 6

Factors that Decrease the Probability of a Research Finding to be True

-
- The smaller the study conducted in a scientific field (i.e., smaller sample size).
 - The smaller the effect sizes in a scientific field.
 - The greater the number of tested relationships and the lesser the selection of tested relationships in a scientific field.
 - The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field.
 - The greater the financial and other interests and prejudices in a scientific field.
 - The hotter a scientific field (with more scientific teams involved).
-

Note. Adapted from Ioannidis (2005).

This paper was published in a medical journal. However, it is very relevant for science in general and psychology as well. The most common approach used in psychological science to declare research findings is Null Hypothesis Significance Testing and p -values. Furthermore, considering the aspects in Table 6, it is known that in psychological research power is low (Sedlmeier & Gigerenzer, 1989), Questionable Research Practices are present (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John, Loewenstein, & Prelec, 2012) and the effect sizes of interested are generally considered to be small or medium (Schäfer & Schwarz, 2019). Thus, the psychological literature needs to be scrutinized. One way by which this is possible, at the level of

single studies, is by replicating the research process that generated those results to see if they hold up again.

2.2 Replicability

What does it mean that a research finding was successfully replicated or did not replicate? To answer these questions, it is important to have a definition of replicability and a measure of replication success. In the literature, there are some definitions of replicability and how it differs from reproducibility. Furthermore, it does not exist a single measure of replication success. Usually, definitions include a combination of considerations such as “consistent effect sizes, consistent distributions, and consistent measures of statistical significance” (Leek & Jager, 2017, p. 26). Other difficulties, such as unknown critical moderators and insufficient power, invite increased prudence when interpreting replication studies’ results.

2.2.1 Replicability and Reproducibility: Definitions

Goodman, Fanelli, & Ioannidis (2016) defined replicability as a particular case of reproducibility. *Reproducibility* is a broad term which can be further distinguished into methods reproducibility, results reproducibility and inferential reproducibility (Figure 5).

Methods reproducibility is what is commonly referred to with just the word *reproducibility*. It means to reach the same conclusions as in the original study, starting from the same dataset and following the same methods and data analysis procedures. This is also called “checking of the analysis” (Tsang & Kwan, 1999).

Results reproducibility corresponds to what we mean with the word *replicability*. It is the achievement of the same results, by following the methodology of the original authors but using a different sample drawn from the same population. If the researcher performs the same procedure, measurement, and analyses, the study can be called a *direct, or exact, replication*. If different measurements and/or manipulations are introduced in the procedure, then the study is a *conceptual replication* (Tsang & Kwan, 1999). The distinction between these two types of replications is

relevant because conceptual replications, even though important, introduce different elements in studies' procedures. This implies that, if a result of a previous study is not successfully replicated in a conceptual replication, the new result can be attributed to the methodological differences between the two studies. On the contrary, direct or exact replications, try to recreate the same conditions of the original study. In this case, if different results are obtained, it is possible to make meaningful comparisons between the two studies' results.

Inferential reproducibility is not so commonly considered, even though it could be the most important among the three. It consists of drawing the same conclusions from a reanalysis of an original study or a replication of a study. This type of reproducibility is different from the previous two because it focuses on the conclusions of a study. In fact, researchers may state different conclusions even from the same analytical results, or there could be a convergence of conclusions starting from different data and studies.

Figure 5. Classification of different types of reproducibility.

2.2.2 Difficulties When Interpreting Replication Studies' Results

Results of large replication projects gave rise to the worrying claim that psychology is facing a replication crisis. Even though concerns are very well justifiable because serious problems exist with the methodology employed in psychological research, conclusions drawn from single replication studies must be cautious. In fact, one replication study (unless having a surprisingly

large sample size) is not enough to declare that an effect is essentially equal to zero, and thus overturn the original finding. Maxwell, Lau, & Howard (2015) warn about the limitations of single replication studies, especially when they are used to sustain the non-existence of a phenomenon. Problems concern statistical power and the sample size needed to prove that an effect is essentially equal to zero.

One issue is how to adequately power a replication study. First, the researcher must answer a difficult question: how much power is, in fact, adequate? Second, the sample size needed to reach a certain level of power depends on the size of the effect under study. If the effect size used to conduct a power analysis for the replication study, corresponds to the one found in the original study, this might be an overestimate due to publication bias and Type M error. As a consequence, the replication study might be underpowered. Third, there exists two types of power: predictive power and conditional power. *Conditional power* is “the probability of rejecting the null hypothesis conditional on an effect size that is presumed to be known with certainty” (Maxwell et al., 2015, p. 489). *Predictive Power* “acknowledges that the effect size is typically not known with certainty but instead is at best an estimate” (Maxwell et al., 2015, p. 489). Predictive power takes the uncertainty of the effect size into account, thus requires bigger sample sizes. All these issues taken together, show that in reality it is difficult to be confident to have adequate power, even in replication studies.

Furthermore, consider a study that has high power (e.g., 80%). This means that if the null hypothesis is false, the study has a high probability of rejecting it. What if this study fails to reject the null hypothesis? Even though this event has a low probability to happen, if H_0 is false, it does not follow that the null hypothesis is actually true. To demonstrate that the null hypothesis of non-existence of an effect is true, other methods are required. In the Frequentist framework, it is possible by establishing a “region of equivalence” where the effect is essentially regarded as equal to zero (Lakens, 2017). In a Bayesian framework, it is possible to establish a region of practical equivalence (ROPE) (Kruschke, 2018). As shown in Maxwell et al. (2015) the sample size required in equivalence studies is very big. For example, 1714 people per group are needed when comparing

two independent groups assuming the same population mean, normality, equal variance and for an equivalence region equal to a Cohens' d of $-0.10 < d < 0.10$ with power = .80 and $\alpha = .05$, two-tailed.

It should come clear that one failed replication with a sample size that is not big enough, is not evidence for the non-existence of a phenomenon. If this is the hypothesis that a researcher wants to test, then multiple studies are needed. Crowd-sourcing projects and Registered Replication Reports using meta-analysis techniques to analyze results are the solutions that have been proposed in the literature to face these issues (Maxwell et al., 2015).

2.3 An Example of Replication Project in Psychology

Since 2012, big replication projects have been conducted with the aim to estimate the replicability of psychological findings. The first one is the Reproducibility Project: Psychology (RP:P) (Open Science Collaboration, 2015). Others are the Many Labs projects (MLP) (Ebersole et al, 2016; Klein et al. 2014; Klein et al., 2018) and the Social Science Replication Project (Camerer et al., 2018).

The Reproducibility Project: Psychology (Open Science Collaboration, 2015), was the first enterprise aiming at estimating the replicability of psychological science and it is also an example of how the word reproducibility had been, improperly, used as a synonym of replicability (Goodman et al., 2016). The teams, involving 270 researchers from 64 different institutions around the world, worked at the replication of 100 studies. These were selected among studies published in three important journals in Psychology in the year 2018 (i.e., Psychological Science, Journal of Personality and Social Psychology and Journal of Experimental Psychology: Learning, Memory and Cognition).

Given that it does not exist a single measure of replication success, results were evaluated using five different indexes. First, the replication effect was evaluated against the null hypothesis of no effect. Ninety-seven percent of the original results were statistically significant, whereas about

36% of the replications produced p -values lower than 0.05. Second, the authors looked if the original effect sizes lied within the 95% confidence interval of the replication study. An overall analysis according to this criterion reported a 47.4% success rate. Third, the original and the replicated effect sizes were compared, showing that 82% of the original effects were larger than the replicated effects. Fourth, the meta-analytic effect was calculated combining the effect sizes of the original and replication studies. Sixty-eight percent of the meta-analytic effects had a confidence interval that did not include zero. Fifth, replication teams gave a subjective assessment of whether the replication was successful or not. Thirty-nine percent of the teams evaluated the replication as successful, producing a similar result to the one obtained using the statistical significance criterion.

A simplistic interpretation of these results might lead to think that the majority of findings in the original studies have been invalidated by the replications. However, as the authors concluded, these replication attempts could not have a final say on whether the original effects were true or false. The results from RP:P demonstrated that there is “room for improvement” in the replicability of Psychology and through replications science can be cumulative, as it should (Open Science Collaboration, 2015, p. 4716-7).

2.4 Questionable Research Practices (QRPs)

The results from replication studies were considered as a red flag for psychological research methods. Furthermore, other considerations on the published literature increased the awareness of possible issues. For example, it is known that statistical power in psychology is low (Sedlmeier & Gigerenzer, 1989). Then, how is it possible that in a sample of studies published between 1986 and 1987, 96% of results were associated with a p -value lower than 5%? (Sterling, Rosenbaum, and Weinkam, 1995). This can be partially explained by the presence of *publication bias* that favors “positive” findings over “null” findings (Franco, Malhotra, & Simonovits, 2014) and by the fact that a low p -value can be obtained exploiting so-called *researchers’ degrees of freedom*. For example, given a dataset, it is possible to perform many different comparisons between groups or

excluding outliers using different techniques or dropping participants with post hoc exclusion criteria (Gelman & Loken, 2014). These are only some examples of Questionable Research Practices that contribute to create findings which upon replication, might not hold up.

P-hacking refers to all those data-analysis pipelines not planned a-priori and generated after looking at the data with the aim of finding a statistically significant result. For instance, it includes adding or removing covariates, introducing new exclusion criteria and performing new analyses on sub-samples (Munafò et al., 2017). This procedure has also been called *data fishing* or *data dredging*. The probability of finding a statistically significant result can be increased by trying out different analyses on a data-set or even by having the possibility of doing so without actually doing it but deciding on the analyses after seeing the data (Gelman & Loken, 2014). However, this process increases also the probability of a finding to be a false positive.

A famous paper by Simmons, Nelson, & Simonsohn (2011) showed how much different researchers' degrees of freedom can inflate the false-positive rate, usually set at 5% (i.e., α). To maintain α at 5%, there is a prerequisite: only one planned analysis is conducted, if more analyses are conducted then α should be adjusted. Consider as an example a between-subjects design where two outcome variables (V_1 , V_2) are measured in two independent samples. In this design, it is possible to conduct three *t*-tests where the null hypotheses are respectively: (1) the population means in V_1 do not differ between the two groups; (2) the population means in V_2 do not differ between the two groups; (3) the population means, in a combined measure of V_1 and V_2 , do not differ between the two groups. If the researchers were to perform all these three tests, setting an alpha level at .05, they would underestimate the number of false positives, which in reality nearly double, rising to 9.5% (Simmons et al., 2011).

Moreover, if only one among those three tests yielded a statistically significant result, and the researchers decided to report only that one test in the final paper, readers won't know that the result was "picked" among other non-significant findings and won't be able to accurately evaluate the validity of those results. This is a Questionable Research Practice called *cherry-picking* and it

generally indicates the decision to report in the paper only those outcome variables that reached statistical significance.

Furthermore, a p -value lower than a certain threshold can be found with 100% success rate if data are peeked repeatedly (Chambers, 2017). In fact, by definition, even when the null hypothesis is true, it is possible to find statistically significant results for a number of times corresponding to the value of α . Furthermore, the p -value is influenced by researchers' intentions, such as when to stop data collection (Wagenmakers, 2007) and p -values are computed under the assumption of independence of observations. This implies that each time data is peeked, α corrections should be applied, or a new experiment should be performed with new observations. Also, stopping rules should be set a priori and not be data-dependent, in fact optional stopping is considered a form of p -hacking.

Even though exploring data to look for possible relationships among variables could shed lights into potentially new interesting research questions, presenting such findings as they have been thought since the planning phase of the study, is a practice that goes against the hypothetico-deductive model of science (Chambers, 2017). In this model, the phase of hypothesis generation comes before data collection and the subsequent data collected are used to test those hypotheses (Figure 6). If unexpected patterns emerge from data, these findings should be reported as exploratory results that will need further confirmation from future studies. Reporting such findings as expected since the beginning and adapting the paper to create a story that fits the data since the introduction section, is a Questionable Research Practice called HARKing, which stands for Hypothesizing After Results are Known (Munafò et al., 2017).

Figure 6. The hypothetical-deductive model of science and threats to this model. Reprinted from ‘A manifesto for reproducible science’, by M.R. Munafò et al., 2017, *Nature Reviews Neuroscience*, 14(5), p.2.

How much widespread are these practices? Table 7 reports the results from two studies conducted among psychologist in the United States (John et al., 2012) and Psychologists in Italy (Agnoli et al., 2017). Practices are listed from the most common to the least common in Italy.

Table 7

Questionable Research Practices in the USA and Italy with Percentage of People Reporting Having Adopted At least One QRP and 95% Confidence Interval in Brackets.

<i>Questionable Research Practice</i>	<i>Italy (Agnoli et al., 2017)</i>	<i>USA (John et al., 2012)</i>
Deciding whether to collect more data after looking to see whether the results were significant.	53.2 (46.6 – 59.7) (<i>N</i> = 222)	55.9 (51.5 – 60.3) (<i>N</i> = 490)
In a paper, failing to report all of a study's dependent measures.	47.9 (41.3 – 54.6) (<i>N</i> = 219)	63.4 (59.1 – 67.7) (<i>N</i> = 486)
In a paper, selectively reporting studies that “worked”.	40.1 (33.6 – 46.6) (<i>N</i> = 217)	45.8 (41.3 – 50.2) (<i>N</i> = 485)
Deciding whether to exclude data after looking at the impact of doing so on the results.	39.7 (33.3 – 46.2) (<i>N</i> = 219)	38.2 (33.9 – 42.6) (<i>N</i> = 484)
In a paper, reporting an unexpected finding as having been predicted from the start.	37.4 (31.0 – 43.9) (<i>N</i> = 219)	27.0 (23.1 – 30.9) (<i>N</i> = 489)
In a paper, “rounding off” a <i>p</i> value (e.g., reporting that a <i>p</i> value of .054 is less than .05).	22.2 (16.7 – 27.7) (<i>N</i> = 221)	22.0 (18.4 – 25.7) (<i>N</i> = 499)
In a paper, failing to report all of a study's conditions.	16.4 (11.5 – 21.4) (<i>N</i> = 219)	27.7 (23.7 – 31.7) (<i>N</i> = 484)
Stopping collecting data earlier than planned because one found the result that one had been looking for.	10.4 (6.4 – 14.4) (<i>N</i> = 221)	15.6 (12.4 – 18.8) (<i>N</i> = 499)
In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do).	3.1 (0.9 – 5.4) (<i>N</i> = 223)	3.0 (1.5 – 4.5) (<i>N</i> = 499)
Falsifying data.	2.3 (0.3 – 4.2) (<i>N</i> = 220)	0.6 (0.0 – 1.3) (<i>N</i> = 495)

Note. Adapted from Agnoli et al. (2017).

2.5 Aren't Scientists Interested in "Finding Things Out"?

Questionable Research Practices are behaviors geared towards results that are positive, novel and clean (Munafò et al., 2017). Should not scientists aim at "finding things out" (Mayo, 2018, p. 7)? That is to say, are not scientists motivated at finding the "true" state of the world and trying as much as possible not to fool themselves?

The reason why researchers might try to seek statistical significance, novelty and good narratives at the cost of reaching false positive results, lies in the incentives system (Smaldino & McElreath, 2016b). Being a researcher is a job on which people make a living on and "publication is the currency of academic science and increases the likelihood of employment, funding, promotion and tenure" (Munafò et al., 2017, p. 7). The characteristics of what reaches publication and what does not, can work as reinforcements for researchers' behavior (Table 8).

Statistical significance is not the only criterion that increases chances of publication, also novelty and aesthetics criteria play a role. Novel findings are preferred because they add new information to the literature, whereas replications of previously published studies are harder to publish (Chambers, 2017). Furthermore, replication studies can be even more hampered by the unavailability of original materials and replicability checks are hard to perform if datasets are not shared or analysis scripts are not available. All these aspects speak against the openness and transparency characteristics which should intuitively be a normal feature of science (Nosek et al., 2015). Finally, aesthetics criteria, which incentivize narrative flows where predictions are confirmed by the data collected, might promote the engagement in HARKing (Giner-Sorolla, 2012).

Table 8

Reinforcements for researchers' behavior

Behavior	Reinforcement
<i>P</i> -hacking statistically significant result and reporting only statistically significant findings (vs. report all pre-planned analyses even though p - value > .05).	Higher chance of publishing because p - value < .05
Perform a new and innovative study (vs. replication of an already published study).	Higher chance of publishing because of novelty characteristics.
Align the initial hypotheses with the results (HARKing) (vs. report all analyses distinguishing confirmatory from exploratory).	Higher chance of publishing because of aesthetics criteria.

2.6 Changing Mindset, Practices and Incentives

The awareness of problems affecting psychological science promoted the creation of new practices to reward transparency and increase incentives to perform replication studies. For example, since 2014 all journals of the Association of Psychological Science offer three badges for studies conducted following some open science practices (i.e., Open Data Badge, Open Materials Badge and Preregistered Badge). This is possible especially thanks to the creation of online repositories, such as the Open Science Framework (OSF) where materials, data, and analyses scripts can be uploaded, identified by a unique URL and made public to the scientific community.

In order to separate more clearly confirmatory analyses from exploratory analyses and prevent *p*-hacking and HARKing, studies can be pre-registered online (e.g., Open Science Framework; Aspredicted.com). The final manuscript of the study can include a link to the pre-registration where the planned hypotheses and data analyses are well-detailed. Another form of pre-registration are Registered Reports (RR). This is an option offered directly by journals where the research process is separated into two stages (Figure 7). In stage 1, researchers submit to the journal the introduction detailing the rationale behind a study, the method section, analyses plan and expected results. A first peer-review process takes place at this stage. If the study is considered

theoretically and methodologically sound, then it is in principle accepted. This means that the authors will conduct the study as detailed in stage 1 and then proceed to stage 2, submitting to the journal the results and discussion sections. At this point, there will be a second round of peer-review mainly aimed at checking if the original plans were respected. If there were no deviations from stage 1, the study is published regardless of the results obtained. If new analyses plans are made after data collection, these can be included in the paper under a section called “exploratory analyses” and clearly distinguished from the a priori plan.

Figure 7. Registered Reports process. Reprinted from ‘Center for Open Science’, 2019, <https://cos.io/rr/>.

Furthermore, in order to incentivize replication studies, some journals offer a particular type of article called Registered Replications Reports (RRR). In this case, the process is the same as a Registered Report however, the study is a replication of a study published in the literature (Simons, Holcombe, & Spellman, 2014). The first Registered Replication Report published in the journal *Psychological Science* looked at an effect in memory research called “Verbal Overshadowing” (Alogna et al., 2014) and will be analyzed in detail in Chapter 4.

Preregistering a study does not only have the benefit to clearly separate exploratory from confirmatory analyses but also to increase researchers’ engagement in the planning phase of a research project and promote the adherence to the hypothetico-deductive model of science (Chambers, 2017). The planning phase of a research project is of fundamental importance, especially from a statistical perspective, for example, to evaluate the feasibility of the study considering the effect size of interest and the sample size needed to detect such an effect. The next chapter will provide a detailed account of a process called *design analysis* that encourages

considerations useful to mindfully plan a research project. The core of a *design analysis* lies in the consideration of a plausible effect size and the risks to which the researcher is exposed in terms of Type I error, Type II error, the factor by which an effect that emerges as statistically significant is overestimated (i.e., Type M error) and the probability that a statistically significant result carries the wrong sign (i.e., Type S error).

Chapter 3

Errors of Magnitude and Sign: Type M and Type S

*“The current replication crisis in science arises
in part from the ill effects of NHST
being used to study small effects with noisy data.
In such settings, apparent success comes easy
but truly replicable results require a more serious
connection between theory, measurement, and data.”*
(A. Gelman, 2018)

In this chapter, we analyze the risks to which researchers are exposed when analyzing results from underpowered studies using thresholds based on statistical significance. In these circumstances, effects that turn out as statistically significant might be exaggerated and/or in the wrong direction. When combining these considerations with the presence of publication bias, it should not be surprising that effects in the literature are biased upwards and that replication studies might observe smaller effect sizes. Researchers can take into account these risks by performing a *design analysis* which allows considering the probabilities of a Type M (magnitude) error and Type S (sign) error in a study. We will define Type M and Type S errors, analyze their relationship with statistical power and conclude the chapter with a practical example of *design analysis*.

3.1 The Statistical Significance Filter in Noisy Studies

Among the many factors contributing to the current “replication crisis” in Psychology and other sciences, a special role has been recognized in the “routine attribution of certainty to noisy data” (Vasishth, Mertzen, Jäger, & Gelman, 2018, p. 151) combined with the statistical significance filter (i.e., the decision criterion on whether to publish a result based on a p -value lower than 0.05) (Loken & Gelman, 2017). Underpowered studies are noisy studies and they seem to be the rule rather than the exception in psychological research (Sedlmeier & Gigerenzer, 1989; Smaldino & McElreath, 2016b). Moreover, results are mainly analyzed based on the statistical significance criterion. As noted by Loken and Gelman (2017) “in noisy research settings, statistical significance

provides very weak evidence for either the sign or the magnitude of the underlying effect” (p. 584). As a consequence, it should not be surprising that replication projects, usually designed to have higher power than the original studies, find results with smaller effect sizes and/or in the opposite direction (Camerer et al., 2018; Open Science Collaboration, 2015).

It is important to note that these considerations are not only exclusive to the use of p -values as thresholds to make decisions. Every time a single statistical index is used to determine publication worthiness with reference to a fixed-line above which a finding is “good” and below which is “bad”, the results in the literature can be severely biased upwards (Amrhein, Korner-Nievergelt, & Roth, 2017; Ioannidis, 2008).

The following paragraphs will first outline what “noise” means and then which risks are associated with statistically significant results in noisy studies, such as the Winner’s Curse and errors of magnitude and sign called Type M error and Type S error (Button et al., 2013; Gelman & Carlin, 2014). This work develops an idea of Gelman and Carlin (2014) with the aim to bring their proposal closer to the psychological context. In fact, these considerations are highly relevant but virtually absent in the current state of research practice.

3.2 What is “Noise”

Noise can be due to sampling variation, measurement error and the influence of extraneous variables on the outcome of interest. *Sampling variation* is inevitable when we are studying samples from a population rather than the entire population. This source of variation is measured by the Standard Error (*SE*) of the test statistic. The standard error is the standard deviation of the sampling distribution of a test statistic. For example, Figure 8 visually represents the standard error of the mean as a test statistic. In this case, the standard error is estimated from the standard deviation of the sample *s*, if the sample size (*N*) is large enough (i.e., at least 30 subjects) (Field, Miles, & Field, 2012):

$$SE = \frac{s}{\sqrt{N}}$$

However, sampling variation can be considered the only source of noise in those cases similar to industrial quality control where measurement is accurate, and sampling happens in a known population. In psychological science, many other sources can generate noise and variation in the data. For instance, *measurement error* which is “the difference between a measured value of quantity and its true value” (Gelman, 2017). This can be particularly problematic when the instruments that are used have not been tested for their validity and reliability and, even more, when measurement scales are built on the fly (Flake & Fried, 2019).

Figure 8. Standard Error of the mean. When conducting a study, only samples from a population of interest are used and not the entire population. If the “real” mean in the population is $\mu = 3$, samples can produce an estimate of the population mean that is slightly different. Imagine taking many samples from the same population and build a distribution with the means found in each sample, the standard deviation of this distribution (i.e., sampling distribution) is the standard error of the sampling distribution of the test statistics. Reprinted from ‘Discovering Statistics Using R’, by A. Field, J. Miles, & Z. Field, 2012, London; Thousand Oaks, Calif: Sage.

How can noise generate statistically significant results? A test statistic is calculated in the sample to check how far it lies from the point value used as the null hypothesis, usually zero. Then, a normal distribution is built around zero to allow for sampling variation, with a standard deviation equal to the standard error. A “statistically significant result” is obtained if the test statistic has a value that lies about two standard errors away from zero. However, the noisier the study, the larger the standard error and the larger the confidence interval associated with the test statistic. A wider confidence interval not only means that the estimate is less precise but also that test statistic can

have larger values falling further away from zero and producing a low p -value. In other words, noise can contribute to find large effect sizes that are statistically significant, because they lie at least about two standard errors away from zero. However, if noise produces an effect, then this effect is not genuine and upon replication, studies will not find that effect again. Figure 9 illustrates this point with an example from simulation studies conducted by Vasishth et al. (2018).

Figure 9. Simulation of repeated sampling where the sample sizes are equal to 20 and 350. Data are generated from a normal distribution with a mean of 15 ms and a standard deviation of 100 ms. The solid horizontal line represents the true population mean. When power is low and findings are statistically significant, the true population mean is overestimated in the correct and in the wrong direction. When power is high, the majority of effects is clustered around the true population mean. Reprinted from ‘The statistical significance filter leads to overoptimistic expectations of replicability’, by S. Vasishtha et. al, 2018, *Journal of Memory and Language*, 103, p.152.

3.3 The Winner's Curse

“High-powered studies have higher *precision* than low-powered studies” (Vasishth et al., 2018, p. 151). In fact, highly powered studies are less noisy because the increased sample size allows decreasing the standard error associated with the estimate. On the contrary, when power is low there exists also greater risk to find a statistically significant effect that exaggerates the real effect size. In these contexts, the apparent win, in terms of obtaining a statistically significant result, is in reality a loss in that “the ‘lucky’ scientist who makes a discovery is cursed by finding an inflated estimate of that effect.” (Button et al., 2013, p. 366). This is called the Winner's curse, and Figure 10 shows an example.

In this hypothetical situation, a researcher is interested in studying an effect that can plausibly be of small dimensions (e.g., Cohen's $d = 0.20$). If he decides to compare two groups on the outcome variable of interest, with 33 people per group (and performing a two-tailed t -test), then he will never be able to simultaneously reject H_0 and find an effect close to what it is considered to be “plausible” (i.e., .20). In fact, 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 rule states that the conclusion should be “fail to reject H_0 ”.

Figure 10. 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. Reprinted from 'Type M & S errors', 2019, by C. Zandonella Callegher, E. Toffalini, G. Bertoldo & Psicostat group, http://147.162.129.70:8788/design_analysis/

Underpowered studies are especially prone to the “Winner’s curse”. However, the concept of power does not make this evident as it is still defined in terms of statistical significance. In fact, power is the probability of finding a statistically significant result, if the null hypothesis is actually false. The problem arises because it is not impossible to find a statistically significant result even if a study is underpowered, and once the null hypothesis has been rejected it might tempting to think only about the probability of the finding to be a false positive (i.e., α) without worrying to interpret the conditions under which it was obtained (i.e., low statistical power).

3.4 Type M and Type S Errors

3.4.1 Definitions

Besides the traditional Type I and Type II errors, Gelman and Carlin (2014) formalized other two types of errors that go beyond a definition of errors based solely on statistical significance and consider also the magnitude and sign of effect sizes. These are called Type M (magnitude) error and Type S (sign) error.

Type M (magnitude) error is also called Exaggeration Ratio. It represents the predictable average overestimation of an effect that emerges as statistically significant. Ideally, a study aims to estimate an effect as accurately as possible (i.e., minimize Type M error). Given that Type M is defined as a ratio between two quantities, the minimum value possible is a value of one and it corresponds to “no exaggeration” of an effect that emerges as statistically significant. The ratio includes at the numerator the mean of all the effect sizes that are statistically significant (in absolute value) and the denominator corresponds to the plausible true effect size.

Type S (sign) error is the probability that a statistically significant result has the opposite sign of the plausible true effect size. If the plausible effect size is positive, then Type S is the ratio between the probability of obtaining a negative statistically significant effect size and the power of the test. If the plausible effect size is negative, then Type S is the ratio between the probability of obtaining a positive statistically significant effect size and the power of the test.

3.4.2 Plausible Effect Size

The cornerstone of the definition of these two types of errors lies in what is a plausible effect size. It might seem paradoxical that researchers have to provide an estimate of the effect size before conducting the experiment, given that they will conduct the study exactly to find what that estimate is. Nevertheless, studies are not usually developed in a void. Instead, hypotheses are derived from theories that make predictions, with a different degree of precision. Strong theories are

those that make precise predictions and could be, in principle, more easily disconfirmed (Kruschke, 2013). For example, a strong theory could predict a medium-sized positive correlation between two variables. In contrast, weak theories make broader predictions, such as predicting that two constructs are correlated without specifying the strength and direction of the correlation (Dienes, 2008). When comparing these two types of theories, the former allows many more observations to disconfirm the theory, whereas the latter allows only the observation of no correlation to disconfirm it. Theories that explain virtually everything and are hard to disconfirm risk to be out of the realm of science (Dienes, 2008). Thus, scientific theories should provide at least a hint about the effect that is expected to be observed.

In practice, it might be undesirable to take as a plausible effect size the estimate found in a pilot study or from a previous study published in the literature. In fact, the plausible effect size refers to what could approximately be the true value of the parameter in the population, whereas the results of single studies or pilots are noisy estimates of that parameter. One possibility could be to take conservative estimates from published studies. Other promising possibilities, though still at the onset, are expert knowledge elicitation or procedures to determine the smallest effect size of interest (SESOI) (Anvari & Lakens, 2019; Button et al., 2013).

3.4.3 Type M, Type S and Statistical Power

As previously mentioned, if a study has low statistical power, a statistically significant result is more likely to be associated with an effect size that exaggerates the real magnitude and has some probability of carrying the wrong sign. As shown by Gelman and Carlin (2014), Type M error and Type S error are related to power (Figure 11). As power decreases, the probability of committing a Type S error or a Type M error increases, even though at different paces. The probability of a Type M error increases exponentially as soon as the power goes below 0.5. Instead, type S error rate is minimized if power is greater than 0.2, increasing exponentially as power becomes lower than 0.2.

Figure 11. Relationship between power, Type S error and Type M error. Reprinted from ‘Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors’, by A. Gelman & J. Carlin, 2014, *Perspectives on Psychological Science*, 9(6), p.644.

Power depends on the true effect size in the population, the sample size in the study and the significance level (i.e., α). As these three factors increase, also statistical power increases. Usually, the element that receives greater attention is sample size. How do Type S error, Type M error and power vary as a function of sample size? Given that power and sample size are directly related, a similar trend is expected. Figure 12 shows an example in the case of a study comparing two different groups on a continuous outcome variable using an independent samples t -test and assuming that the plausible effect size corresponds to a Cohen’s d of 0.35 and establishing an α of 0.05. It can be noticed that with increasing sample size, power increases reaching 80% with about 130 people per group. With this sample size, the exaggeration ratio is approximately 1 and, as the sample size decreases, Type M error increases reaching approximately a value of 2 with 30 participants per group. This means that if a statistically significant result is obtained with this sample size, the effect size is expected to be on average two times the plausible effect size. Finally, the probability of finding a statistically significant result in the wrong direction becomes worrisome with less than 50 participants per group.

Figure 12. Power, Type M error and Type S error as a function of sample size in an independent samples *t*-test, assuming a plausible effect size equal to a Cohen's *d* of 0.35 and α of 0.05. Reprinted from 'Enhancing statistical inference in psychological research via prospective and retrospective design analysis.' by G. Altoè et. al, 2019.

This example also illustrates an interesting point. Many times, 80% is recommended as a desirable level of power without any specific reason. Even though a magic number that fits every situation does not exist and postulating one could invite mindless decision-making, this example shows that with a power of 0.80 the probability of finding a statistically significant effect in the wrong direction is minimized (i.e., Type S approximates zero) and, on average, the effect size will be exaggerated by only about 10% (Figure 13).

Figure 13. Relationship between power, Type S error and Type M error. When power reaches 80%, Type M and Type S are minimized.

3.5 Design Analysis: Prospective and Retrospective

Gelman & Carlin (2014) suggested using Type S and Type M errors in what is called a *design analysis*, which can be seen as a broader definition of a power analysis. With increasing interest in studies preregistration, researchers are more commonly performing prospective power analysis to estimate the necessary sample size to detect the effect of interest. Similarly, a *prospective design analysis* is conducted in the planning phase of a research project in order to choose a sample size that allows not only to maximize power but also to minimize the probabilities of committing a Type M error and Type S error.

The same analysis can also be performed once a study has already been conducted and results are in hand and it is called *retrospective design analysis*. In this analysis, the plausible effect size, and not the effect size found in the study, is used to analyze the power of the study, Type M error and Type S error. This provides information about the extent to which the results of the study could be exaggerated and/or in the wrong direction aiding the interpretation of statistically significant results.

Even though both procedures can be of great value, *prospective design analysis* is much more useful if the aim is to maximize the informativeness of studies and allocate resources in projects that are meaningful and reasonable. Instead, a *retrospective design analysis* is valuable in those cases where it has not been feasible to recruit enough subjects to reach a certain level of power, or when a study is underpowered because of reasons of other nature. In these cases, knowing what the power of a study, Type M and Type S errors are, can help to avoid overoptimistic interpretation of results.

3.5.1 Case Study

In the following paragraphs, a case of prospective and retrospective *design analysis* is illustrated using as an example Study 2 in the paper “A functional basis structure-seeking: Exposure to structure promotes willingness to engage in motivated action” (Kay, Laurin, Fitzsimons, & Landau, 2014). To illustrate a *prospective design analysis* in practice, the reader will be asked to imagine that the study has not been yet conducted and that researchers were to perform this analysis when planning the study. Further on, the reader will be asked to imagine that researchers had to stop data collection earlier because of unforeseen circumstances (e.g., lack of funding) and decided to perform a *retrospective design analysis* on the data collected.

The theoretical basis for the studies presented in Kay et al. (2014) arose 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 perceiving a structured world could increase people’s willingness to make efforts and sacrifices towards their own goals. The authors decided to test this hypothesis using a between-subjects design where participants were randomly assigned to two different conditions that differed in the type of text they had to read (i.e., random vs. structure). 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. Figure 14 illustrates the procedure of the study. Before being allocated to one of the two conditions, participants were asked to write down a long-term goal deemed important and towards which they could make progress on a daily basis. After reading the text, their willingness to work towards that goal was measured. 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.

Figure 14. Procedure in Study 2 of Kay et al. (2014).

3.5.2 Prospective Design Analysis

Imagine that before conducting the study, the group of researchers would like to pre-register their hypotheses and estimate the feasibility of the research project in terms of the number of people they will need to recruit. Therefore, they decide to conduct a *prospective design analysis*. At this stage, they face an important and difficult point: establishing what is the size of the effect that they expect to be “plausible”. That is to say, what is the difference in willingness to pursue goal-directed behaviors that can plausibly be observed between the two groups? As previously mentioned, these considerations could be based on published findings by taking a conservative estimate because of publication bias and exaggeration of effect sizes in the literature. Also, they could base the estimate on their experience as experts on the topic and/or other experts’ suggestions. For the sake of this example, it is assumed that a plausible effect size corresponds to a Cohen’s d of 0.20.

Table 9

Function “design_analysis ()”

Type of design analysis	Input	Output
Prospective	<ul style="list-style-type: none"> ○ Plausible effect size (Cohen’s d) ○ Significance level α ○ Power 	<ul style="list-style-type: none"> ○ Sample size per group ○ Type M Error ○ Type S Error
Retrospective	<ul style="list-style-type: none"> ○ Plausible effect size (Cohen’s d) ○ Significance level α ○ Sample size per group 	<ul style="list-style-type: none"> ○ Power ○ Type M Error ○ Type S Error

Notes. Significance level α is set at 0.05 by default but can the user can change it. Groups must have the same sample size.

In order to perform a *prospective design analysis*, they use the function “design_analysis ()” (Table 9)¹. This is similar to the usual functions employed for prospective power analysis where three out of four information (i.e., sample size per group, plausible effect size expressed as Cohen’s d , power, and significance level α) are given as input and the fourth is returned. However, in the “design_analysis” function the plausible effect size has always to be specified. Furthermore, the significance level α is set by default at 0.05 but can be changed by the user. If the user would like to perform a *prospective design analysis*, then s/he has to specify the desired level of power and the plausible effect size. Subsequently, the function returns the sample size per group and the associated Type M error and Type S errors. Instead, if the user would like to perform a *retrospective design analysis*, then s/he has to specify the sample size used and the plausible effect size. The function returns the associated levels of power, Type M error and Type S error. This allows the researchers to explore different scenarios in terms of probability of finding a statistically significant result, if an effect actually exists (i.e., power) and also to make considerations regarding the associated risks in terms of magnitude of the effect sizes and the associated sign, if a result were to turn out as statistically significant.

Table 10 illustrates results from different scenarios in the current example. Type S Error does not seem to be a relevant issue in all four scenarios whereas the exaggeration rate ranges from

¹ Code available in Appendix A.

30% to 5% when going from a 60% power to 90% power. However, such an increase in power and decrease in Type M error, requires more than double the sample size. We assume that after some discussions, the researchers deem reasonable to recruit 784 participants which would give them 80% power and a Type M Error Rate of 1.13.

Table 10

Prospective Design Analysis for Cohen's $d = 0.20$ ²

Cohen's d	Power	Sample Size (per group)	Sample Size (total)	Type M Error	Type S Error
0.2	0.90	532	1064	1.05	0
	0.80	392	784	1.13	0
	0.70	314	628	1.2	0
	0.60	244	488	1.3	0

3.5.3 Retrospective Design Analysis

Imagine that after starting the study, the researchers realize not having enough funding to continue recruiting participants and they choose to stop the study. At this point, they recruited only 68 people (34 per group)³ and they decide to analyze the data collected. An independent samples t -test is performed to test for the difference between the means of the two groups on the measure of goal-directed behavior ($M_{structure} = 5.26$, $SD_{structure} = 0.88$; $M_{random} = 4.72$, $SD_{random} = 1.32$). Results seem promising: $t(66) = 1.98$, $p = .05$, Cohen's $d = 0.51$ ⁴. If they were to stop interpreting results at this point following the usual “Null Ritual”, they would reject the null hypothesis of no difference between the means of the two groups, and maybe even go further and accept the alternative hypothesis that there exists a difference between the two groups and that perceiving the world as structured rather than random could increase the willingness to work towards personal goals. Furthermore, this effect seems to be of medium size (i.e., Cohen's $d = 0.51$), giving further evidence for the presence of an effect.

² Code available in Appendix A.2

³ The original study, Kay et al. (2014), only reported the total sample size of 67 people without specifying the number of participants per group. In the current example, an even number of participants per group is hypothesized, leading to a total sample of 68 participants. The reason behind this decision was that the `design_analysis()` function requires equal sample sizes.

⁴ Calculations for the current example are reported in Appendix B.

However, the researchers decide to perform a *retrospective design analysis* on these results, especially because they did not expect such a large effect size at the beginning of the study when they made various considerations on what could be a plausible effect size in their experiment. Table 11 shows the results of the *retrospective design analysis* considering as a plausible effect size a Cohen's d of 0.20. In this case, only the plausible effect size and the actual sample size are given as input. It is possible to note that with this sample size, the statistical power is about 13%. This means that if an effect exists, there is only 13% probability of finding a statistically significant result. Researchers would fall in the “What does not kill statistical significance makes it stronger fallacy” if they were to conclude that, despite the low probability of finding a statistically significant result, they were able to find $p = 0.5$, thus their result is even more remarkable. This is clearly understandable when also Type M Error and Type S Error probabilities are considered. In fact, in such a design the predictable average overestimation of an effect that emerges as statistically significant is 3.04 (i.e., on average, the effect is overestimated by 204%) and the probability of finding a statistically significant result in the wrong direction is 2%.

Table 11

Retrospective Design Analysis for a Cohen's $d = 0.20$ and a Total Sample Size of 68 peoples

Cohen's d	Sample Size (per group)	Sample Size (total)	Power	Type M Error	Type S Error
0.20	34	68	0.13	3.04	0.02

Note that in a *retrospective design analysis*, it is used the plausible effect size (e.g., Cohen's $d = 0.20$) and not the effect size found in the study (e.g., Cohen's $d = 0.50$). Using the effect size found in the study would mean to perform what is called “post-hoc power analysis”, which is renown in the literature to be problematic (Gelman, 2019). In fact, results of single studies are noisy estimates of the parameter in the population and do not represent the plausibly true parameter. Furthermore, this procedure could be used to support a circular reasoning in which obtaining a

⁵ Code available in Appendix A.2

small effect size implies lower power in the study and obtaining a large effect size leads to a higher power in the study.

In conclusion, this situation is the one depicted in Figure 10 illustrating the Winner's curse. If the researchers did not perform such *retrospective design analysis* and would solely interpret results based on the significant p -value, they would conclude that there is a medium difference between the two groups (Cohen's d of .50) without realizing that, even assuming that a difference exists, the results of this study are providing an effect that is too big to be true.

Study 2 in Kay et al. (2014) has been object of a replication in the “Many Labs 2 project” (Klein et al., 2018) with a large sample size ($n = 6506$). Results did not support the original findings: $t(6498.63) = -0.94.98$, $p = .35$, Cohen's $d = -0.02$, $95\% CI = [-0.07, 0.03]$. Interested readers are invited to look at the response from the original authors to the replication project (Laurin et al., 2018).

Chapter 4

Type M Error Might Explain the Decline Effect

*“It is generally very difficult to prove
that something does not exist;
it is considerably easier to show
that a tool is inadequate
for studying that something.”
(U. Simonsohn, 2015)*

In this chapter, we present various explanations that have been hypothesized to analyze the decline in the magnitude of effect sizes in replication studies. We propose to use a *retrospective design analysis* to explore the possibility that the original effect was exaggerated and to gain insights about the adequacy of the original study to investigate the phenomenon of interest. If the Type M error rate in the original study was high, then highly powered replications of that study will have a higher probability to observe a decline effect. We illustrate such a proposal with a case study analyzing two experiments on the Verbal Overshadowing effect (Schooler & Engstler-Schooler, 1990) and their replications (Alogna et al., 2014).

4.1 The Decline Effect: Definition and Types

One way to evaluate replication studies' results, is to compare the effect size obtained in the replication to the one in the original study. It has been noticed that “the strength of evidence for a particular hypothesis usually declines over time, with replication studies presenting smaller effects than original studies” (Amrhein et al., 2017, p.11). For instance, the Open Science Framework (2015) reported that “overall, original study effect sizes ($M = 0.403$, $SD = 0.188$) were reliably larger than replication effect sizes ($M = 0.197$, $SD = 0.257$)” (p. 4716-5) and in the Social Science Replication Project (Camerer et al., 2018) “the effect size of the replication is on average about 50% of the original effect size”.

The term *decline effect* has been introduced in the literature to indicate “the notion that science routinely observes effect sizes decreasing over repeated replications for reasons that are still not well understood” (Schooler, 2014, p. 579). As discussed in chapter 3, when underpowered studies are evaluated using the statistical significance filter, effects might be exaggerated. As a consequence, when higher powered replication studies are conducted, there is some probability that the effects will be smaller and/or in the opposite direction. Thus, Type M errors and Type S errors in the original studies might partially explain the decline effect.

However, statistical power cannot be the only explanation of the decline effect. Protzko & Schooler (2017) proposed a classification of different types and explanations for the decline effect, as well as general mechanisms that play a role in these different situations (Table 12). Even though the explanations are not mutually exclusive, it is nevertheless important to understand which sources are behind each type of decline effect to improve the design of studies and the reliability of research findings.

Table 12

Mechanisms and Types of Decline Effects (Protzko & Schooler, 2017)

<i>General mechanisms</i>	
<i>Underpowered studies</i>	The original study was underpowered and in underpowered studies statistically significant effect sizes are commonly exaggerated. Since replication studies are designed to assure higher power, a decline effect emerges.
<i>Publication bias</i>	Studies with statistically significant results are written-up and published more often than studies with non-significant findings. As a consequence, “null findings” appear later, if ever, in the literature producing a distort picture of the effects.
<i>Selective reporting</i>	Various forms of Questionable Research Practices (QRPs) twist results so that there is a higher chance of publication. QRPs can produce results which are not genuine and might partially contribute to observe decline effects in replication studies.
<i>Types - Explanations</i>	
<i>False positive decline effect</i>	There was never an effect in the first place, thus with subsequent replications, there is a regression towards the true null mean ⁶ .
<i>Inflated decline effects</i>	The effect exists however, the original study overestimated the effect size.
<i>Under-specified decline effects</i>	Some factors (moderators) that were critical to obtain the effect were absent in the replication. Once these relevant factors are recognized, a comparable effect size is found.
<i>Genuinely decreasing decline effects</i>	Some effects naturally decline, for example, because of changes in the population (e.g., cultural developments and prejudices).
<i>Indeterminate and non-conventional decline effects</i>	The mere observation of a phenomenon contributes to the decline of the phenomenon under study.

⁶ The reverse case, that the replication is a false negative and the original study found a genuine effect, is not reported here given that the authors proposing the decline effect did not present this case, even though it could be a possible scenario.

In this chapter, the focus is on the connection between *underpowered studies* and the *inflated decline effect* type. Once an exact replication finds a different effect size compared to the original study, a *retrospective design analysis* could help to understand what the Type M error rate and the statistical power of the original study were. This could be done by using as a plausible effect size a range of effect sizes (e.g., small, medium and large), as well as the effect size found in the replication in line with the small telescope approach to evaluate replication results (Simonsohn, 2015).

According to the small telescope approach, once a replication of a study is conducted, it is possible to analyze the design of the original study to evaluate if it would have been able to detect an effect of the size found in the replication. The telescope analogy goes as follow: “imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one” (Simonsohn, 2015, p. 560). It is possible to combine the small telescope approach with the *design analysis* approach whereby the effect size found in the replication can be used to study the power, Type M error and Type S error probabilities of the original study.

In the following paragraph, we will illustrate a case study on how a *retrospective design analysis* can provide insights into the type of decline effect. The analysis will look at two studies (Study 1 and Study 4) published in the paper “Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid” (Schooler & Engstler-Schooler, 1990). These studies are interesting for various reasons. First of all, they have been the object of a multisite replication project involving 31 laboratories in different countries (Alogna et al., 2014). Furthermore, this replication project was the first-ever published Registered Replication Report in the journal Psychological Science (Simons et al., 2014) and received the collaboration of the original author, Johnathan Schooler, who also proposed the aforementioned classification of decline effects. This study also had great media

attention appearing in “The New Yorker” in the article: “The Truth Wears Off: Is there something wrong with the scientific method?” (Lehrer, 2010). Finally, the Verbal Overshadowing Effect suggests that, contrary to what is usually found on studies looking at verbal rehearsal and memory, verbally describing a visual memory, impairs the recognition of that image, instead of increasing the accuracy. This is extremely relevant because of the real-world implications that these findings might have for recommendations on how to conduct lineup procedures and for the evaluation of eyewitness evidence in criminal trials (Chin, 2014).

4.2 Case Study

4.2.1 The Original Study

In 1990 Schooler and Engstler-Schooler published an article suggesting that “when to be remembered materials are faces, verbal rehearsal hurts rather than helps memory performance” (Alogna et al., 2014, p. 557). This conclusion was suggested from the results of a series of studies. In Study 1 (Figure 15), participants watched a video of a bank robbery and then completed a filler task (i.e., crossword puzzle) for 20 minutes. Afterward, half of the participants had to write down the description of the perpetrator for 5 minutes (experimental group), whereas the other half of participants had to write down capitals of the United States (control group). Finally, both groups were presented with a photo lineup containing eight people where they could either pick a person who they thought to be the perpetrator or do not choose anybody. Given that the lineup contained the picture of the actual perpetrator, each participant could either make a correct identification or one of two types of errors: identifying the wrong person (i.e., misidentification) or not choosing anyone (i.e., miss). In Study 4, participants performed the very same tasks with only two differences: order and timing of the filler task. In Study 4 (Figure 16), the description of the perpetrator or listing of capitals was performed after watching the video, whereas in Study 1 it was done before the lineup. Moreover, the filler task lasted 10 minutes whereas in Study 1 it lasted 20 minutes.

Figure 15. Procedure of Study 1 in Schooler and Engstler-Schooler (1990) and in the Registered Replication Report 2. Reprinted from ‘Registered Replication Report: Schooler and Engstler-Schooler (1990)’, by Alogna et al., 2014, *Perspectives on Psychological Science*, 9(5), p.558.

Figure 16. Procedure of Study 4 in Schooler and Engstler-Schooler (1990) (S&E-S Study4) and in the Registered Replication Report 1 (RRR1). Reprinted from ‘Registered Replication Report: Schooler and Engstler-Schooler (1990)’, by Alogna et al., 2014, *Perspectives on Psychological Science*, 9(5), p.558.

Two replication studies were conducted, one for Study 1 and one for Study 4. The replication of Study 1 is an exact replication in which the procedure followed exactly the original (Figure 15). The replication of Study 4 is not an exact replication (Figure 16) in that the duration of the filler task was different from the original (i.e., 10 minutes) to the replication (i.e., 20 minutes). In the history of the replication project (Alogna et al., 2014), replication of Study 4 was performed before, so it is called Registered Replication Report 1 (RRR1), and replication of Study 1 started after RRR1, but importantly before these results were known, thus it is called Registered Replication Report 2 (RRR2).

4.2.2 Study 4 and the Registered Replication Report 1

The data and results of the original study (Study 4) are reported in Table 13 and they show that “subjects in the Face Verbalization condition were significantly less accurate at identifying the target face than Control subjects (with respective accuracy rates of 49 and 71%) $\chi^2(1) = 3.92, p < .05$ ” (Schooler & Engstler-Schooler, 1990, p. 54). The risk ratio, calculated as the difference in accurate identifications between the verbal description group and the control group, was equal to -22% , 95% $CI[-44\%, -1\%]$. Hence, participants in the verbal description group showed a reduced “risk” of correctly identifying the robber compared to participants who engaged in a task unrelated to the crime scene.

Table 13

Data Study 4 in Schooler & Engstler-Schooler (1990)

	Correct Identification	Incorrect Identification or Miss	Total
Verbal description Condition	18	19	37
Control Condition	27	11	38
Total	45	30	75

In the replication study RRR1, 3729 participants were recruited from 31 laboratories across the world. Results are reported in Table 14 where each cell presents the sum of participants falling in that condition across all laboratories (Appendix D). The meta-analytic effect size was a risk ratio of -4.01% , 95% $CI[-7.15\%, -0.87\%]$, showing a marked decrease from the original Study 4. Furthermore, the replication effect size is not contained in the 95% confidence interval of the original study (Figure 17). As reported by the authors of the replication: “it is unclear whether the effect actually declined in size or whether the original estimate was just an inaccurate estimate of the effect” (Alogna et al., 2014, p. 570). It is hard to test if there was a real decline over time. However, it is possible to perform a *retrospective design analysis* to evaluate the accuracy of the estimate in the original study.

Table 14

Data Registered Replication Report (1) in Alogna et al. (2014)

	Correct Identification	Incorrect Identification or Miss	Total
Verbal description Condition	950	923	1873
Control Condition	1014	842	1856
Total	1964	1765	3729

Figure 17. Risk Ratios with 95% confidence interval for the two original studies and their corresponding replications. RRR1 was the replication of Study 4. RRR2 was the replication of Study 1. Code available in Appendix C.

Some preliminary computational considerations are necessary to perform the *retrospective design analysis* in this case. The function “design_analysis” requires the plausible effect size to be expressed in terms of Cohen’s d and to be positive. Even though this information is not directly present in the replication’s paper (Alogna et al., 2014), it is possible to calculate these values starting from the contingency tables of the data collected (Table 13; Table 14). These transformations are reported in Appendix D. In the original study, the Verbal Overshadowing effect was equal to a Cohen’s d of 0.56 whereas, in the replication, Cohen’s d was equal to 0.09.

At this point, it is possible to perform a *retrospective design analysis* hypothesizing four different scenarios: the plausible effect size being equal to a small effect size (Cohen's $d = 0.20$), medium effect size (Cohen's $d = 0.50$), large effect size (Cohen's $d = 0.80$), and the effect size found in the replication (Cohen's $d = 0.09$). Table 15 shows the retrospective statistical power, the Type M and Type S errors in these four scenarios⁷.

Table 15

*Retrospective Design Analysis of Study 4 in Schooler & Engstler-Schooler (1990)*⁸

Sample size (total)	Hypothetical effect size (Cohen's d)	Power	Type M	Type S
76	0.2	0.14	2.92	0.02
	0.5	0.58	1.33	0
	0.8	0.93	1.05	0
	0.09	0.07	6.17	0.16

Considering that the original study found a statistically significant effect, with such a design (i.e., total sample size of 76 participants) only hypothesizing that the true effect was large (Cohen's $d = 0.80$) the researchers would have minimally overestimated the effect size (i.e., Type M error rate of 1.05). If the true effect size were to be medium (Cohen's $d = 0.50$), the effect would have been on average overestimate by 33%. Finally, if the real effect size were to be small (Cohen's $d = 0.20$), the overestimation of the effect would greatly increase, reaching a Type M error rate of 2.92 with a probability of 2% of the effect being in the wrong direction. When considering as the plausible effect size the one found in the replication (Cohen's $d = 0.09$), the original study had an extremely low power (i.e., 7%) with an associated high probability of exaggerating the magnitude of the effect and/or finding an effect in the wrong direction as indicated by the Type M error rate of 6.17 and the Type S probability of 16%. These considerations point to the direction that the original

⁷ The function `design_analysis` requires equal sample size per group, whereas Study 4 had 37 participants in the Verbal Description Condition and 38 participants in the Control Condition. For the retrospective design analysis an equal number of 38 participants per group was set, resulting in a total of 76 participants.

⁸ Code available in Appendix A.2

study might have been severely underpowered and by finding a statistically significant effect size the effect could have been an overestimation. Thus, this could be a case of *inflated decline effect*.

As previously described, the replication (RRR1) is slightly different from the original Study 4 because of the duration of the filler task (i.e., 20 minutes in the replication vs. 10 minutes in the original). If the time lapse between the phase of verbal description and the recognition task is a relevant factor, then this would be a case of both *under-specified decline effect* and *inflated decline effect*. In fact, if timing is an important moderator of the verbal overshadowing effect and given the procedural difference between the two studies, the different time-lapses could be partially responsible for the observed decline.

However, for the practical applications of these findings outside the laboratory, it is important to note that the effect size observed in the replication is very small (Cohen's $d = 0.09$). Hence, researchers should evaluate if this difference is relevant for practical purposes or if it should be considered practically equivalent to zero. This analysis could be performed both in a Bayesian and a Frequentist framework (Kruschke, 2018; Lakens, 2017). Furthermore, if timing between the verbal description and the lineup identification is relevant, it should be noted that usually there exist a difference of days or weeks between the time when the eyewitness reports the description of the perpetrator and the lineup administration. Thus, the Verbal Overshadowing Effect could be even more attenuated in real-life contexts (Schooler, 2014).

Finally, could the original study have been a false positive? A false positive means that even though the effect is non-existent (e.g., the accuracy of people who provide a verbal description is exactly the same as those who do not), the researcher is able to reject the null hypothesis of no difference and declare that a difference exists. Given that researchers set α at 5%, in repeated replications of the same experiment, this event should happen no more than 5% of the times. However, it has been noticed that rarely the phenomena of interest in the social and behavioral sciences are exactly equal to zero and one of the weaknesses of the NHST approach is that it will be almost always possible to reject the null hypothesis of no-effect because exactly null effects are

rare but they could be, at best, extremely small (Gelman & Carlin, 2014). However, if the evaluation of results shifts from solely rejecting null hypotheses to considering measures of effect sizes combined with the associated Type M and Type S errors probabilities, it would be clear when effect sizes are statistically significant but too small to be of practical importance or too big to be true.

In conclusion, similar considerations were made by Simonsohn (2015) using the small telescope approach. He suggested that even though the replication was consistent with the original study (i.e., both studies found an effect in the same direction), the design of the original study could be considered inadequate given that it would have never been able to detect the effect found with the bigger telescope (i.e., the replication study).

4.2.3 Study 1 and the Replication Registered Report 2

Given the procedural differences between the replication study RRR1 and Study 4, 22 out of the 31 laboratories collaborated also to perform an exact replication of Study 1 in Schooler & Engstler-Schooler (1990) (RRR2). In the original study, it was found that “Verbalizing the previously seen face substantially reduced recognition performance. The target face was correctly picked by 38% of subjects in the Face Verbalization condition and 64% of Control subjects, $\chi^2(1) = 5.93, p < .05$ ” (Schooler & Engstler-Schooler, 1990, p. 44). The data of the study are shown in Table 16 and the risk ratio was calculated as the difference in accurate identifications between the verbal description group and the control group (Risk ratio = -25% , $95\% CI[-45\%, -5\%]$). Thus, the accuracy of the identification was worst in the group that verbally described the robber compared to the control group.

Table 16

Data Study 1 in Schooler & Engstler-Schooler (1990)

	Correct Identification	Incorrect Identification or Miss	Total
Verbal description Condition	17	27	45
Control Condition	28	16	43
Total	44	44	88

In the replication study RRR2, the total sample size was equal to 2055 and results are reported in Table 17. The meta-analytic effect size was a risk ratio of -16.31% , $95\% CI[-20.47\%, -12.14\%]$, showing a decrease from the original Study 1. However, the replication effect size was within the 95% confidence interval of the original study (Figure 17). One way to analyze the possibility of an *inflated decline effect* (original Cohen's $d = 0.56$ vs. replication Cohen's $d = 0.36$) is by performing a *retrospective design analysis* in the four different scenarios: Cohen's $d = 0.20$, Cohen's $d = 0.50$, Cohen's $d = 0.80$ and the effect size found in the replication (Cohen's $d = 0.36$). Transformations from risk ratio to Cohen's d are reported in Appendix D.

Table 17

Data Registered Replication Report (2) in Alogna et al. (2014)

	Correct Identification	Incorrect Identification or Miss	Total
Verbal description Condition	394	635	1029
Control Condition	559	467	1026
Total	953	1102	2055

Table 18

*Retrospective Design Analysis of Study 1 in Schooler & Engstler-Schooler (1990)*⁹

Sample size (total)	Hypothetical effect size (Cohen's d)	Power	Type M	Type S
88	0.2	0.15	2.67	0.01
	0.5	0.64	1.26	0
	0.8	0.96	1.03	0
	0.36	0.39	1.6	0

Results of the *retrospective design analysis* are shown in Table 18. The original study found a statistically significant effect but with such a design (i.e., total sample size of 88 participants), only hypothesizing that the true effect was large (Cohen's $d = 0.80$) the researchers would have minimally overestimated the effect size (i.e., Type M error rate of 1.03). If the true effect size were to be medium (Cohen's $d = 0.50$), the effect would have been on average overestimated by 26%. Finally, if the real effect size were to be small (Cohen's $d = 0.20$), the overestimation of the effect would reach a Type M error rate of 2.67 with a probability of 1% of the effect being in the wrong direction. In conclusion, no decline effect would be observed in a replication study only if the real effect size were to be equal to a Cohen's d of about 0.80. However, even the original study found an effect size of 0.56, which indicates that probably this expectation is not realistic. Instead, even when hypothesizing that the real effect is 0.50, with such a design a statistically significant result would nevertheless on average overestimate the effect of almost 30%, thus a small decline effect would be expected upon a large replication. In conclusion, applying the “telescope” approach (Simonsohn, 2015) and considering as a plausible effect size the one found in the Registered Replication Report (Cohen's $d = 0.36$), the original study was not adequate to study the effect of interest. In fact, with a total sample size of 88 people the study has 39% power and a statistically significant result is expected to be associated with an effect size that is on average exaggerated by 60%.

⁹ Code available in Appendix A.2

Chapter 5

Conclusions

The crisis of confidence in published research findings in psychological science had some beneficial side-effects. The intrinsic motivation of scientists to find out how phenomena really work is leading the inspection of scientific practices. The aim is not only to declare the current state of the art but to increasingly refine scientific standards and work towards the improvement of Psychology as a science. Therefore, what has been called so far “Psychology’s Crisis” can be better rephrased as “Psychology’s Renaissance” (Nelson, Simmons, & Simonsohn, 2018). New practices are spreading, for instance, studies’ pre-registration, multi-site collaboration, platforms to share data and materials in an open science perspective and increased awareness that mindful statistical practice is necessary when conducting research.

The present work aimed at providing a small contribution analyzing one of the most cited issues: the seemingly never-ending problem of underpowered studies. In particular, the focus has been on the evaluation of results originating from noisy studies and their role in the replication crisis. Undue confidence in such findings might surprise researchers witnessing the decline of effects in replication studies where noise is more contained.

Considering the probability of committing a Type M and a Type S error can improve statistical inferences operating both at the stage of research design and during results’ evaluation. As presented in the third chapter, if the effects of interest are small, the risks associated with underpowered studies are serious: statistically significant results can overestimate the real effect and can have the wrong sign. Then, in a publication system that rewards statistically significant findings, effect sizes in the literature might be unreliable. These considerations have highly applied value, especially when these studies are used to advise real-world interventions or for the lay public that is not acquainted with the limitations of research methods. Studies with high Type M and Type S error probabilities might suggest that one intervention is more effective than another when the

reality is the other way around or there is no difference at all. Therefore, when the aim is exclusively to find statistically significant results, it is easy to lose sight of what researchers are really interested to know, which is the magnitude and direction of effects.

It should be noted that in the present work we used a very simple case (i.e., independent samples *t*-test) whereas researchers constantly face problems that cannot be captured by such a simple model. This is a limitation which can be overcome by future implementations of *design analysis* in more complex scenarios. However, here the main aim was to show the potentials of *design analysis* to change the mindset with which a research endeavor is approached. In particular, thinking hard about effect sizes could help to shift from the black and white thinking associated with the quest for statistically significant results and increase the interest in estimation of effect sizes with precision. Another limitation of the current work is that the plausible effect size is formalized as a point value. This might not accurately reflect researchers' expectations which can instead be better captured by an interval of values with an associated probability distribution. This suggestion is currently under development in the work of Altoè et al. (2019). Finally, in the current work, we chose Cohen's *d* as an effect size measure to illustrate a *design analysis* because it is an effect size measure widely used in psychology. However, the code (Appendix A.1) could be easily adapted to implement a design analysis also with other effect size measures. This was one of the main reasons behind the choice to use resampling methods to estimate power, Type M and Type S errors instead of using an analytical approach.

As mentioned in the first chapter, statistics offers a toolbox that includes also Bayesian statistics (Gigerenzer & Marewski, 2015). Even though it might be tempting to think that leaving behind the Frequentist framework would solve the issue of effect sizes' exaggeration, this problem remains relevant in a Bayesian framework as well. In particular, the use of Bayes Factors for hypothesis testing is becoming increasingly popular. However, changing statistical analysis does not change the fact that calculations are based on noisy studies looking at small effects. When this is coupled with the use of thresholds to declare evidence for one hypothesis or the other

(Schönbrodt & Wagenmakers, 2018), then the effect sizes in these studies might be biased. Thus, the extensions of Type M and Type S errors in a Bayesian framework remains an interesting development for future work.

In conclusion, the main benefit of performing a *design analysis* is to clearly put forward considerations about plausible effect sizes and analyze results based on that. Establishing what is a plausible effect size is the central, and possibly most difficult, point of the process. However, in a discipline that aims at using quantitative tools, it should be taken at heart the precision of estimates and measurements issues, such as the reliability and validity of the instruments that are used. In the words of Gelman (2015) on noisy studies looking at small effect sizes: “My best analogy is that they are trying to use a bathroom scale to weigh a feather—and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down”.

References

- Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, 12(3), e0172792. <https://doi.org/10.1371/journal.pone.0172792>
- Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9(5), 556–578. <https://doi.org/10.1177/1745691614545653>
- Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagni, A., Finos, L., & Pastore, M., (2019). *Enhancing statistical inference in psychological research via prospective and retrospective design analysis*. Manuscript submitted for publication.
- Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *PeerJ*, 5, e3544. <https://doi.org/10.7717/peerj.3544>
- Anvari, F., & Lakens, D. (2019, February 1). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. <https://doi.org/10.31234/osf.io/syp5a>
- Bishop, D. V. M. (2018). *Fallibility in Science: Responding to Errors in the Work of Oneself and Others*. 7.
- Borenstein, M. (Ed.). (2009). *Introduction to meta-analysis*. Chichester, U.K: John Wiley & Sons.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (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., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science

References

- between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637.
<https://doi.org/10.1038/s41562-018-0399-z>
- Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Chin, J. M. (2014). Psychological science's replicability crisis and what it means for science in the courtroom. *Psychology, Public Policy, and Law*, 20(3), 225–238.
<https://doi.org/10.1037/law0000012>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates)
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London ; Thousand Oaks, Calif: Sage.
- Fisher, R. A. (1936). Design of experiments. *Br Med J*, 1(3923), 554-554.
- Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. <https://doi.org/10.31234/osf.io/hs7wm>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.
<https://doi.org/10.1126/science.1255484>
- Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management. Advance Online Publication*.

- Gelman, A. (2015, April 21). The feather, the bathroom scale, and the kangaroo [Blog post]. Retrieved from <https://statmodeling.stat.columbia.edu/2015/04/21/feather-bathroom-scale-kangaroo/>
- Gelman, A. (2016, June 26). When are people gonna realize their studies are dead on arrival? [Blog post]. Retrieved from <https://statmodeling.stat.columbia.edu/2016/06/26/29449/>
- Gelman, A. (2017, February 11). Measurement error and the replication crisis [Blog post]. Retrieved from <https://statmodeling.stat.columbia.edu/2017/02/11/measurement-error-replication-crisis/>
- Gelman, A. (2018). Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size. *Annals of surgery*, 20, 1.
- 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. (2019, February 27). Statistical-significance filtering is a noise amplifier. [Blog post]. Retrieved from <https://statmodeling.stat.columbia.edu/2019/02/27/statistical-significance-filtering-is-a-noise-amplifier/>
- 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., & Carlin, J. (2017). Some Natural Solutions to the p -Value Communication Problem—And Why They Won't Work. *Journal of the American Statistical Association*, 112(519), 899–901. <https://doi.org/10.1080/01621459.2017.1311263>
- Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, 102(6), 460. <https://doi.org/10.1511/2014.111.460>
- Gigerenzer, G. (1993). The Superego, the Ego, and the Id in Statistical Reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences*. Lawrence Erlbaum Associates.

References

- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606.
<https://doi.org/10.1016/j.socec.2004.09.033>
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409).
<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.
- Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science*, 7(6), 562–571. <https://doi.org/10.1177/1745691612457576>
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps12–341ps12.
<https://doi.org/10.1126/scitranslmed.aaf5027>
- 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.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, 7(1), 1–20.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124.
- Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology*, 19(5), 640–648. <https://doi.org/10.1097/EDE.0b013e31818131e7>
- 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>

- 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>
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A “many labs” replication project. *Journal of Open Psychology Data*, 2(1).
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... 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>
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270-280.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4), 355-362.
- Laurin, K., Kay, A. C., and Landau, M. J. (2018). Structure and goal pursuit: Individual and cultural differences. *Advances in Methods and Practices in Psychological Science* 1, 491–494. [doi:10.1177/2515245918797130](https://doi.org/10.1177/2515245918797130)

References

- Leek, J. T., & Jager, L. R. (2017). Is Most Published Research Really False? *Annual Review of Statistics and Its Application*, 4(1), 109–122. <https://doi.org/10.1146/annurev-statistics-060116-054104>
- Lehrer, J. (2010, December 13). The Truth Wears Off: Is there something wrong with the scientific method? Retrieved November 8, 2018, from The New Yorker website:
<https://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off>
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585. <https://doi.org/10.1126/science.aal3618>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70(6), 487–498. <https://doi.org/10.1037/a0039400>
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). <https://doi.org/10.1017/9781107286184>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology*, 69, 511-534.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Oakes, M. W. (1986). *Statistical inference*. Epidemiology Resources.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. <https://doi.org/10.1126/science.aac4716>

- Protzo, J. & Schooler, J. W. (2017). Decline Effects: Types, Mechanisms, and Personal Reflections. In S. O. Lilienfeld & I. W. Waldman (Eds.), *Psychological Science Under Scrutiny* (pp.85-107). John Wiley & Sons, Inc.
- Royall, R. (2017). *Statistical evidence: a likelihood paradigm*. Routledge.
- Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology, 10*. <https://doi.org/10.3389/fpsyg.2019.00813>
- Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review, 25*(1), 128-142.
- Schooler, J. W. (2014). *Turning the Lens of Science on Itself: Verbal Overshadowing, Replication, and Metascience*.
- Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*(1), 36–71. [https://doi.org/10.1016/0010-0285\(90\)90003-M](https://doi.org/10.1016/0010-0285(90)90003-M)
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at *Perspectives on Psychological Science. Perspectives on Psychological Science, 9*(5), 552–555. <https://doi.org/10.1177/1745691614543974>
- Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science, 26*(5), 559–569. <https://doi.org/10.1177/0956797614567341>
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science, 3*(9), 160384.

References

- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49(1), 108-112.
- Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and Theory Development in Organizational Science: A Critical Realist Perspective. *The Academy of Management Review*, 24(4), 759. <https://doi.org/10.2307/259353>
- 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>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>

Appendices

Appendix A: R functions for *Design Analysis*¹⁰ and Examples

1. Function

```
# Clear workspace
rm(list=ls())

#
require("truncnorm")

## Loading required package: truncnorm

#

##### It calculates Cohen's d [Internal Function]
cohen_d <- function(x1,x2) {

  (m1 <- mean(x1, na.rm = TRUE))
  (m2 <- mean(x2, na.rm = TRUE))
  S1 <- sum((x1-m1)^2, na.rm=TRUE)
  S2 <- sum((x2-m2)^2, na.rm=TRUE)
  (n1 <- length(na.omit(x1)))
  (n2 <- length(na.omit(x2)))
  (df <- n1+n2-2)

  (pool.s <- sqrt( (S1 + S2) / df ) )
  (d <- (m2 - m1)/pool.s)

  s1 <- sd(x1, na.rm = TRUE)
  s2 <- sd(x2, na.rm = TRUE)
  (tstat <- (m2-m1) / (pool.s * sqrt((n1+n2)/(n1*n2))))
  (pval <- pt(abs(tstat),df,lower.tail=FALSE)*2)

  return(list(d=d,tstat=tstat,pval=pval))
}

##### It calculates power, typeM error & typeS error based on B replicates [Internal Function]
# Note for design_analysis: target_d is equal to sim_y
sim_estimate <- function(sim_y, target_d, n1 = n1, n2 = n1, sig.level = .05, B
= 1e4 ) {
```

¹⁰ This function was developed within the Psicostat group (<http://ip146179.psy.unipd.it/psicostat/web/>), with special contributions from Altoè Gianmarco, Bertoldo Giulia, Pastore Massimiliano, Toffalini Enrico, Zandonella Callegher Claudio.

```

outsim <- t(replicate( B, {
  x1 <- rnorm(n1)
  x2 <- rnorm(n2,sim_y)
  d <- cohen_d(x1,x2)
  sim <- c(d$d,d$tstat,d$pval,sim_y)
}))
colnames(outsim) <- c("est_d","tstat","pval","sim_d")
outsim <- data.frame(outsim)
(outsim$power <- sum( outsim$pval < sig.level ) / nrow(outsim))
(outsim$typeS <- with(outsim, sum( (pval < sig.level) & (tstat<0)) / sum(pval
< sig.level) ) )
(outsim$typeM <- with(outsim, mean(abs(est_d[pval< sig.level])))) / target_d)
return(outsim)
}

##### It performs retrospective design analysis [Internal Function]
design_analysis_power <- function( n, d, sig.level = 0.05, B = 1e4 ,tol=.005 )
{
  out <- sim_estimate(sim_y=d,target_d=d,n1=n,sig.level=sig.level,B=B)
  out1=list(d=d,n=n,power=out[1,"power"],typeS=out[1,"typeS"],typeM=out[1,"type
M"])
  return(out1)
}

##### It performs prospective design analysis [Internal Function]
design_analysis_n <- function( d, power, sig.level = 0.05, rangen = c(2,1000),
B = 1e4, tol = .005 ){

  n_seq <- seq( rangen[1], rangen[2], by = 1 )
  (n_target <- round(median(n_seq)))
  find_power <- FALSE

  ## check with maximum N
  #cat("Estimating power with n =",rangen[2],"\\n")
  (est_P <- sim_estimate( n1 = rangen[2], n2 = rangen[2], sim_y = d, target_d=d
, sig.level = sig.level, B = B ))
  (est_power <- est_P[1,]$power)

  if ( est_power < power ) {
    cat(paste0("Actual power = ", est_power, " with n = ", rangen[2], " (per gr
oup); " ),"\\n")
    cat(paste0(" try to increase maximum of rangen > ", rangen[2],"."),"\\n")
    out <- NULL
  } else {

    ## estimating power
    while( (!find_power) ) {
      #cat("Estimating power with n =",n_target,"\\n")

```

```

    (est_P <- sim_estimate( n1 = n_target, n2 = n_target, sim_y = d, target_d
=d, sig.level = sig.level, B = B))
    (est_power <- est_P[1,]$power)

    if ( (est_power<=(power+tol)) & (est_power>(power-tol)) ) {
      find_power <- TRUE
    } else {

      if (length(n_seq)==1) {
        print(n_seq)
        stop(" ")
      }

      if ( est_power > (power-tol) ) {
        (n_seq <- seq( min(n_seq), n_target, by = 1))
        (n_target <- round(median(n_seq)))
      } else {
        (n_seq <- seq( n_target, max(n_seq), by = 1))
        (n_target <- round(median(n_seq)))
      }
    }
  }
  out <- list( d = d, power = power, n = n_target, typeS=est_P[1,"typeS"],typeM=est_P[1,"typeM"])
}

#cat("", "\n")
if (!is.null(out)) return(out)
}

##### It performs prospective and retrospective design analysis
# Note: plausible 'd' is a fixed value. 'n' is the sample size per group
design_analysis <- function( d, n = NULL, power = NULL, sig.level = 0.05,
                           B = 1e4 , rangen = c( 2, 1000 ), tol=.005) {

  if (d <= 0){
    stop("A d greater than 0 must be entered")
  }

  if (sum(sapply(list(n, d, power, sig.level), is.null)) != 1) {
    stop("It is necessary to provide either 'n' or 'power'")
  }

  if (is.null(n)) {
    out <- design_analysis_n( d=d, power=power, sig.level=sig.level, rangen=rangen, B=B )
  } else {

```

```

    if (is.null(power)) {
      out <- design_analysis_power( n, d, sig.level, B )
    }
  }

  if (!is.null(out)) return(out)
}

##### Sampling Cohen's d given a bounded distribution [Internal Function]
sampling_d <- function( target_limits = c(0,1), distribution = c("uniform","normal"), k = 1/6, B0 = 1e4 ) {

  distribution <- match.arg(distribution)
  my <- (sum(target_limits)/2)
  sy <- diff(target_limits)*k
  if (distribution=="uniform") {
    y <- runif(B0,target_limits[1], target_limits[2])
  } else {
    y <- rtruncnorm(B0,target_limits[1], target_limits[2], mean = my, sd = sy)
  }
  return(list(y=y,my=my,sy=sy))
}

##### It performs retrospective design analysis given a plausible interval and a distribution for d.
##### It works also for different sample sizes per group.
##### It works also given a fixed value of d.
design_est<- function( n1 , n2 = n1, target_d = NULL, target_d_limits = NULL, distribution = c("uniform","normal"),
                      k = 1/6, sig.level = 0.05, B = 500, B0 = 500, return_data = FALSE ) {

  # CHECKS
  distribution <- match.arg(distribution)

  if (sum(sapply(list(target_d,target_d_limits), is.null)) != 1) {
    stop("It is necessary to provide either 'target_d' or 'target_d_limits'.")
  }

  if (!is.null(target_d)) {
    if (target_d<=0) stop("Target_d must be greater than 0")
  }

  if (!is.null(target_d_limits)) {
    if (min(target_d_limits)<0) stop("Minimum of target_d_limits must be greater than 0")
  }
}

```

```

    if (distribution=="uniform") k<-NULL
  }

  call_fix <- list(n1=n1, n2=n2, target_d=target_d, B=B)
  call_est <- list(n1=n1, n2=n2, target_d_limits=target_d_limits, distribution=
distribution,k=k,B=B,B0=B0)

##### plausible interval and distribution for d
if(!is.null(target_d_limits))
{
  my <- (sum(target_d_limits)/2)
  sy <- diff(target_d_limits)*k
  y <- sampling_d(target_d_limits,distribution = distribution,B0=B0, k = k) #
##
  sim_cohen <- y$y
  #
  data <- t(sapply(1:B0, function(b){
    temp <- sim_estimate(sim_cohen[b], target_d=my, n1 = n1, n2 = n2, B = B )
    c(temp[1,"power"],temp[1,"typeS"],temp[1,"typeM"])
  })))
  data <- data.frame(data)
  data<-data.frame(data)
  names(data)<-c("power","typeS","typeM")

  (power <- mean(data$power))
  (typeS <- mean(data$typeS,na.rm = TRUE))
  (typeM <- mean(data$typeM,na.rm = TRUE))

  if (return_data) {
    results <- list(power=power,typeS=typeS,typeM=typeM,data=data)
    outList <- list(call=call_est,results=results)
  } else {
    results <- list(power=power,typeS=typeS,typeM=typeM)
    outList <- list(call=call_est,results=results)
  }
  return(outList)
}
##### fixed d
if (is.null(target_d_limits)) {
  temp<-sim_estimate(sim_y=target_d, target_d=target_d, n1 = n1, n2 = n2, B =
B )
  results<-list(power=temp[1,"power"],typeS=temp[1,"typeS"],typeM=temp[1,"typ
eM"])

```

```

    outList <- list(call=call_fix,results=results)
    return(outList)
  }
}

```

2. Examples¹¹

Prospective Design Analysis - Table 10 (Chapter 3)

```
design_analysis(d = 0.20, power = 0.90)
```

```
## $d
## [1] 0.2
##
## $power
## [1] 0.9
##
## $n
## [1] 532
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.057436

```

```
design_analysis(d = 0.20, power = 0.80)
```

```
## $d
## [1] 0.2
##
## $power
## [1] 0.8
##
## $n
## [1] 392
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.135142

```

```
design_analysis(d = 0.20, power = 0.70)
```

¹¹ Results might slightly vary because the function uses resampling methods to estimate power, Type S and Type M errors. The number of iterations is now set at 10000. However, it is possible to increase the iterations to obtain more stable results.

```

## $d
## [1] 0.2
##
## $power
## [1] 0.7
##
## $n
## [1] 314
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.198263

design_analysis(d = 0.20, power = 0.60)

## $d
## [1] 0.2
##
## $power
## [1] 0.6
##
## $n
## [1] 244
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.296718

# Retrospective Design Analysis - Table 11 (Chapter 3)
design_analysis(n = 34, d = 0.20)

## $d
## [1] 0.2
##
## $n
## [1] 34
##
## $power
## [1] 0.1344
##
## $typeS
## [1] 0.02232143
##
## $typeM
## [1] 3.044466

```


Retrospective Design Analysis of Study 4 - Table 15 (Chapter 4)

design_analysis (n = 38, d = 0.2)

```
## $d
## [1] 0.2
##
## $n
## [1] 38
##
## $power
## [1] 0.1408
##
## $typeS
## [1] 0.015625
##
## $typeM
## [1] 2.922419
```

design_analysis (n = 38, d = 0.5)

```
## $d
## [1] 0.5
##
## $n
## [1] 38
##
## $power
## [1] 0.5714
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.330982
```

design_analysis (n = 38, d = 0.8)

```
## $d
## [1] 0.8
##
## $n
## [1] 38
##
## $power
## [1] 0.9299
##
## $typeS
## [1] 0
##
```

```

## $typeM
## [1] 1.051293

design_analysis (n = 38, d = 0.09)

## $d
## [1] 0.09
##
## $n
## [1] 38
##
## $power
## [1] 0.0683
##
## $typeS
## [1] 0.1581259
##
## $typeM
## [1] 6.159292

# Retrospective Design Analysis of Study 1 - Table 18 (Chapter 4)
design_analysis (n = 44, d = 0.2)

## $d
## [1] 0.2
##
## $n
## [1] 44
##
## $power
## [1] 0.1528
##
## $typeS
## [1] 0.01112565
##
## $typeM
## [1] 2.672299

design_analysis (n = 44, d = 0.5)

## $d
## [1] 0.5
##
## $n
## [1] 44
##
## $power
## [1] 0.6412
##
## $typeS

```

Appendices

```
## [1] 0
##
## $typeM
## [1] 1.26302

design_analysis (n = 44, d = 0.8)

## $d
## [1] 0.8
##
## $n
## [1] 44
##
## $power
## [1] 0.9633
##
## $typeS
## [1] 0
##
## $typeM
## [1] 1.032417

design_analysis (n = 44, d = 0.36)

## $d
## [1] 0.36
##
## $n
## [1] 44
##
## $power
## [1] 0.381
##
## $typeS
## [1] 0.0007874016
##
## $typeM
## [1] 1.604505
```

Appendix B: Results Case Study Chapter 3

```

# Results of the study considering 68 participants (34 per group)

# Mean of goal-directed measure in "structure" condition
m1 <- 5.26

# Standard deviation of goal-directed measure in "structure" condition
s1 <- 0.88

# Mean of goal-directed measure in "random" condition
m2 <- 4.72

# Standard deviation of goal-directed measure in "random" condition
s2 <- 1.32

# Sample size per group
n = 34

# t-test
(t_test <- ((m1 - m2)/(sqrt(((s1^2)/n)+((s2^2)/n)))))

## [1] 1.984764

# Degrees of freedom
(df <- n+n-2)

## [1] 66

# p-value
(pval <- pt(abs(t_test),df,lower.tail=FALSE)*2)

## [1] 0.0513301

# Pooled standard deviation
(pool_sd <- sqrt( (s1 + s2) / 2 ) )

## [1] 1.048809

# Cohen's d
(d <- (m1 - m2)/pool_sd)

## [1] 0.5148698

```

Appendix C: Code Figure 17

```
# Figure 17: Risk Ratios with 95% Confidence Intervals
(RR_visual <- matrix (c("Original - Study 4", "Replication - RRR1", "Original -
Study 1", "Replication - RRR2", -0.22, -0.0401, -0.25, -0.1631, -0.1, -0.0087,
-0.05, -0.1214, -0.44, -0.0715, -0.45, -0.2047), nrow = 4))

##      [,1]      [,2]      [,3]      [,4]
## [1,] "Original - Study 4" "-0.22"  "-0.1"   "-0.44"
## [2,] "Replication - RR1"  "-0.0401" "-0.0087" "-0.0715"
## [3,] "Original - Study 1" "-0.25"  "-0.05"   "-0.45"
## [4,] "Replication - RR2"  "-0.1631" "-0.1214" "-0.2047"

dimnames(RR_visual) <- list("Rows" = c("Study 4", "RRR1", "Study 1", "RRR2"), "
Columns"=c("Study", "RR", "CI_low", "CI_high"))
RR_visual <- as.data.frame(RR_visual)
RR_visual

##           Study      RR  CI_low CI_high
## Study 4 Original - Study 4  -0.22   -0.1  -0.44
## RRR1      Replication - RR1 -0.0401 -0.0087 -0.0715
## Study 1 Original - Study 1  -0.25   -0.05  -0.45
## RR2       Replication - RR2 -0.1631 -0.1214 -0.2047

RR_visual$RR <- as.numeric(as.character(RR_visual$RR))
RR_visual$CI_low <- as.numeric(as.character(RR_visual$CI_low))
RR_visual$CI_high <- as.numeric(as.character(RR_visual$CI_high))
str(RR_visual)

## 'data.frame':    4 obs. of  4 variables:
## $ Study : Factor w/ 4 levels "Original - Study 1",...: 2 3 1 4
## ..- attr(*, "names")= chr "Study 4" "RRR1" "Study 1" "RR2"
## $ RR : num -0.22 -0.0401 -0.25 -0.1631
## $ CI_low : num -0.1 -0.0087 -0.05 -0.1214
## $ CI_high: num -0.44 -0.0715 -0.45 -0.2047

library(ggplot2)
ggplot(RR_visual, aes(x=Study, y=RR)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_errorbar(aes(ymax = CI_high, ymin = CI_low), size = .5, height = 0.2, co
lor ="gray50") +
  geom_point(size = 1.5, color ="#4169E1") +
  theme_bw() +
  labs(y="Risk Ratio", x = "Study") +
  coord_flip()
```

Appendix D: Transformation to Cohen's d

Cohen's d can be derived from a contingency table where the Odds Ratio (OR) have been calculated. The formula is presented in Borenstein (2009):

$$\text{Cohen's } d = \frac{\ln(OR) \times \sqrt{3}}{\pi}$$

1. Study 4

```
# Study 4: From Risk Ratio to Cohen's d
# Contingency table
(study4_vd_correct <- 18)
## [1] 18

(study4_vd_incorrect <- (37-18))
## [1] 19

(study4_control_correct <- 27)
## [1] 27

(study4_control_incorrect <- (38-27))
## [1] 11

study4 <- matrix(c(study4_vd_correct,study4_control_correct,study4_vd_incorrect
,study4_control_incorrect),ncol=2)
colnames(study4) <- c("Correct Identification","Incorrect Identification or Miss
s")
rownames(study4) <- c("Verbal Description", "Control Condition")
(study4 <- as.table(study4))

##              Correct Identification Incorrect Identification or Miss
## Verbal Description             18                      19
## Control Condition             27                      11
#Totals
(study4_vd <- study4_vd_correct + study4_vd_incorrect)
## [1] 37

(study4_control <- study4_control_correct + study4_control_incorrect)
## [1] 38

(study4_total <- study4_vd + study4_control)
## [1] 75

# Odds Ratio on the contingency table (odds of making a correct identification
in the control condition over the odds of making a correct identification in th
```

```

e verbal description condition)
library(mosaic)
(OR_study4 <- oddsRatio(study4, verbose = TRUE))

##
## Odds Ratio
##
## Proportions
##   Prop. 1:  0.4865
##   Prop. 2:  0.7105
##   Rel. Risk: 1.461
##
## Odds
##   Odds 1:  0.9474
##   Odds 2:  2.455
##   Odds Ratio: 2.591
##
## 95 percent confidence interval:
##   0.9905 < RR < 2.154
##   0.9996 < OR < 6.716
## NULL
## [1] 2.590909

# Transformation of Odds Ratio to Log Odds Ratio
(logOR_study4 <- log(OR_study4))
## [1] 0.9520088

# Cohen's d
(d_study4 <- logOR_study4*sqrt(3)/(pi))
## [1] 0.52487

# Replication - RR1: From Risk Ratio to Cohen's d
# Contingency table
(r1_vd_correct <- 32 + 37 + 17 + 34 + 32 + 29 + 40 + 15 + 40 + 24 + 33 + 26 + 4
5 + 26 + 34 + 12 + 24 + 36 + 23 + 28 + 40 + 34 + 27 + 28 + 34 + 20 + 37 + 49 +
32 + 26 + 36)
## [1] 950

(r1_vd_incorrect <- (6 + 7 + 14 + 18 + 22 + 23 + 9 + 9 + 15 + 11 + 12 + 8 + 31
+ 19 + 11 + 21 + 12 + 17 + 11 + 11 + 4 + 15 + 7 + 15 + 6 + 21 + 7 + 20 + 13 + 1
4 + 11) + (18 + 12 + 20 + 18 + 21 + 22 + 10 + 17 + 26 + 15 + 18 + 14 + 34 + 17
+ 17 + 18 + 19 + 15 + 21 + 11 + 11 + 11 + 16 + 8 + 15 + 11 + 8 + 18 + 11 + 11 +
20))
## [1] 923

(r1_control_correct <- 40 + 33 + 18 + 27 + 42 + 41 + 40 + 24 + 39 + 28 + 39 + 2
2 + 58 + 35 + 26 + 20 + 31 + 23 + 23 + 21 + 38 + 39 + 32 + 26 + 31 + 28 + 37 +
55 + 30 + 30 + 38)

```

```
## [1] 1014

(r1_control_incorrect <- (10 + 13 + 19 + 32 + 19 + 20 + 13 + 15 + 14 + 12 + 15 +
16 + 35 + 15 + 9 + 17 + 12 + 18 + 18 + 18 + 7 + 4 + 6 + 23 + 7 + 17 + 10 + 30 +
11 + 12 + 16) + (7 + 4 + 13 + 11 + 14 + 15 + 8 + 7 + 16 + 10 + 14 + 12 + 17 + 1
1 + 16 + 13 + 12 + 9 + 14 + 11 + 10 + 17 + 13 + 12 + 15 + 5 + 4 + 13 + 10 + 12
+ 14))
## [1] 842

RR1 <- matrix(c(r1_vd_correct, r1_control_correct, r1_vd_incorrect, r1_control_incorrect), ncol=2)
colnames(RR1) <- c("Correct Identification", "Incorrect Identification or Miss")
rownames(RR1) <- c("Verbal Description", "Control Condition")
(RR1 <- as.table(RR1))

##              Correct Identification Incorrect Identification or Miss
## Verbal Description                950                      923
## Control Condition                1014                      842

#Totals
(r1_vd <- r1_vd_correct + r1_vd_incorrect)
## [1] 1873

(r1_control <- r1_control_correct + r1_control_incorrect)
## [1] 1856

(r1_total <- r1_vd + r1_control)
## [1] 3729

# Odds Ratio on the contingency table (odds of making a correct identification
in the control condition over the odds of making a correct identification in the
verbal description condition)
(OR_RR1 <- oddsRatio(RR1, verbose = TRUE))

##
## Odds Ratio
##
## Proportions
##      Prop. 1:  0.5072
##      Prop. 2:  0.5463
##      Rel. Risk: 1.077
##
## Odds
##      Odds 1:  1.029
##      Odds 2:  1.204
##      Odds Ratio: 1.17
##
## 95 percent confidence interval:
```


Appendices

```
## 1.013 < RR < 1.145
## 1.029 < OR < 1.331
## NULL
## [1] 1.170049

# Transformation of Odds Ratio to Log Odds Ratio
(logOR_RR1 <- log(OR_RR1))
## [1] 0.1570454

# Cohen's d
(d_RR1 <- logOR_RR1*sqrt(3)/(pi))
## [1] 0.08658368
```

2. Study 1

```
# Study 1: From Risk Ratio to Cohen's d
# Contingency table
(study1_vd_correct <- 17)
## [1] 17

(study1_vd_incorrect <- (16+11))
## [1] 27

(study1_control_correct <- 28)
## [1] 28

(study1_control_incorrect <- (10+6))
## [1] 16

study1 <- matrix(c(study1_vd_correct, study1_control_correct, study1_vd_incorrect,
  study1_control_incorrect), ncol=2)
colnames(study1) <- c("Correct Identification", "Incorrect Identification or Miss")
rownames(study1) <- c("Verbal Description", "Control Condition")
(study1 <- as.table(study1))

##              Correct Identification Incorrect Identification or Miss
## Verbal Description                17                      27
## Control Condition                 28                      16

#Totals
(study1_vd <- study1_vd_correct + study1_vd_incorrect)
## [1] 44

(study1_control <- study1_control_correct + study1_control_incorrect)
## [1] 44
```

```

(study1_total <- study1_vd + study1_control)
## [1] 88

# Odds Ratio on the contingency table (odds of making a correct identification
in the control condition over the odds of making a correct identification in th
e verbal description condition)
(OR_study1 <- oddsRatio(study1, verbose = TRUE))

##
## Odds Ratio
##
## Proportions
##      Prop. 1:  0.3864
##      Prop. 2:  0.6364
##      Rel. Risk: 1.647
##
## Odds
##      Odds 1:  0.6296
##      Odds 2:  1.75
##      Odds Ratio: 2.779
##
## 95 percent confidence interval:
##      1.067 < RR < 2.543
##      1.172 < OR < 6.591
## NULL
## [1] 2.779412

# Transformation of Odds Ratio to Log Odds Ratio
(logOR_study1 <- log(OR_study1))
## [1] 1.022239

# Cohen's d
(d_study1 <- logOR_study1*sqrt(3)/(pi))
## [1] 0.5635901

# Replication - RR2: From Risk Ratio to Cohen's d
# Contingency table

(r2_vd_correct <- 25 + 27 + 8 + 23 + 32 + 26 + 13 + 15 + 15 + 15 + 23 + 11 + 15
+ 13 + 13 + 7 + 13 + 16 + 11 + 20 + 27 + 26)
## [1] 394

(r2_vd_incorrect <- (7 + 13 + 13 + 7 + 26 + 25 + 10 + 8 + 8 + 7 + 7 + 9 + 13 +
7 + 13 + 10 + 13 + 19 + 10 + 5 + 8 + 19) + (18 + 13 + 9 + 20 + 17 + 21 + 23 + 2
0 + 20 + 8 + 19 + 10 + 22 + 21 + 20 + 13 + 18 + 17 + 29 + 13 + 12 + 15))
## [1] 635

```

Appendices

```
(r2_control_correct <- 31 + 36 + 16 + 24 + 50 + 38 + 26 + 26 + 15 + 19 + 24 + 1
4 + 26 + 19 + 22 + 14 + 24 + 17 + 23 + 24 + 35 + 36)
## [1] 559

(r2_control_incorrect <- (15 + 10 + 7 + 17 + 12 + 13 + 11 + 11 + 17 + 8 + 13 + 8
+ 8 + 8 + 11 + 12 + 9 + 20 + 13 + 6 + 7 + 19) + (4 + 8 + 8 + 9 + 13 + 16 + 11 +
9 + 9 + 3 + 13 + 8 + 16 + 14 + 13 + 4 + 8 + 13 + 14 + 9 + 5 + 5))
## [1] 467

RR2 <- matrix(c(r2_vd_correct, r2_control_correct, r2_vd_incorrect, r2_control_inc
orrect), 2)
colnames(RR2) <- c("Correct Identification", "Incorrect Identification or Miss")
rownames(RR2) <- c("Verbal Description", "Control Condition")
(RR2 <- as.table(RR2))

##              Correct Identification Incorrect Identification or Miss
## Verbal Description                394                      635
## Control Condition                559                      467

#Totals
(r2_vd <- r2_vd_correct + r2_vd_incorrect)
## [1] 1029

(r2_control <- r2_control_correct + r2_control_incorrect)
## [1] 1026

(r2_total <- r2_vd + r2_control)
## [1] 2055

# Odds Ratio on the contingency table (odds of making a correct identification
in the control condition over the odds of making a correct identification in th
e verbal description condition)
(OR_RR2 <- oddsRatio(RR2, verbose = TRUE))

##
## Odds Ratio
##
## Proportions
##   Prop. 1:  0.3829
##   Prop. 2:  0.5448
##   Rel. Risk: 1.423
##
## Odds
##   Odds 1:  0.6205
##   Odds 2:  1.197
##   Odds Ratio: 1.929
##
## 95 percent confidence interval:
```

```
## 1.293 < RR < 1.566
## 1.618 < OR < 2.3
## NULL
## [1] 1.929179

# Transformation of Odds Ratio to Log Odds Ratio
(logOR_RR2 <- log(OR_RR2))
## [1] 0.6570943

# Cohen's d
(d_RR2 <- logOR_RR2*sqrt(3)/(pi))
## [1] 0.362275
```