UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Magistrale in
Psicologia Clinica dello Sviluppo

Tesi di Laurea Magistrale

Do Registered Reports Improve Sample Size Planning in Psychology? An exploratory study

I Registered Reports hanno migliorato la pianificazione della numerosità campionaria nella ricerca psicologica? Uno studio esplorativo

Relatore:
Prof. Gianmarco Altoè

Laureanda: Rebecca Norsa
Matricola: 2016879

Anno Accademico 2021/2022

To my brother

Thanks

I would like to express my deepest appreciation to Prof. Altoè, the completion of my dissertation would not have been possible without his guidance and nurturing.

I would also like to extend my sincere thanks to Claudio Zandonella Callegher, for his valuable advice, assistance and patience.

I cannot leave University of Padua without mentioning my beloved friends: Chythia Alvarez Moreno, Ludovica Bonsignore, Mikael Poli, Veronica Ganeo for their unwavering support.

I am extremely grateful to my parents, their profound belief in my work and in my abilities.

Table of Contents

Summary

Credibility crisis in scientific literature began almost two decades ago, in 2005, when Ioannidis (2005) published the innovative article *Why most published research findings are false*. From 2005 onward, many have been the attempt to improve replicability in order to reconquer credibility. Over the last years, researchers replicated a considerable number of original studies published in the literature, with the aim to test the trustworthiness of the findings. Replication studies' results were not terribly encouraging. In fact, a worrisome number of replications found effects which were not as strong as in the original studies or they did not find any effect at all, underlying both methodological and publishing issues. The present work aims to analyze the replicability crisis in psychology focusing on Power Analysis detection, with specific attention towards sample size planning. The main objective is to highlight whether Registered Reports enhances sample size planning, with a specific focus on psychology. Therefore, the current work aims to revise psychological inherent Registered Reports (RRs) published on Open Science Framework (OSF), as of January 2022.

This dissertation is divided into five chapters.

The first chapter chronicles the roots of the credibility crisis, how it happened, and why reproducibility and replicability are vital to debunk the crisis.

The second chapter introduces Null Hypothesis Significance Testing (NHST) which is the prevalent approach used for statistical inference in the social sciences. Since NHST is a mix of Fisher and Neyman-Pearson frameworks, these two approaches will

also be included in the chapter. Furthermore, key elements of power analysis and design analysis involved in sample size planning are presented.

The third chapter narrates the threats of publication process; on this point, Questionable Research Practices (QRP) will be largely describes. A section is dedicated to Registered Reports, a form of pre registration introduces as a solution to cope with Questionable Research Practices.

The fourth chapter summarizes the information gathered by analyzing Registered Reports published on Open Science Framework (OSF) as of January 2022; for this purpose, we built a dataset in order to proceed with an exploratory study.

Finally, the last chapter aims to analytically discuss the findings.

The current work has been written in Rmarkdown.

Chapter 1

Reproducibility and Replicability

In this chapter, we give an overview of what we mean by psychology credibility crisis. Initially, we present the origins of the crisis of confidence in the published literature in psychological. We discuss the importance to rely on Statistical Inference to run trustworthy researches. Afterwards, we bring up the essential value of reproducibility and replicability in scientific progress. We then point out difficulties when interpreting results, indeed conclusions obtained from single replication studies must be careful. Later, we analyze positive consequence of the crisis which lead to the reward of transparency and an incentive to replicate studies This chapter concludes with a brief presentation of the aim of this dissertation.

1.1 Credibility Crisis

During the last years many fields faced an *out of ordinary credibility crisis*: science, biology, medicine in addition to every statistics dependent fields. Typically, 2005 is indicated as the beginning of the crisis, when Ioannidis (2005) published the innovative article *Why most published research findings are false*, which might be considered the starting point of the increased awareness of issues with published literature, therefore the start of the *crisis of confidence*.

Ioannidis (2005) paper was published in a medical journal, nevertheless it involved every statistics related field, including psychology.

The aforementioned article emphasized that a large number of research's findings are unclear, unreliable or cannot be replicated. The author strongly stated the presence of ill-founded strategy of claiming conclusive research on the base of a single study or assessed by formal statistic significance, typically for p-value less than 0.05 (Ioannidis, 2005).

On this point, Ioannidis (2005) identified as one of the chief reasons that explains the high rate of un-replicability of research the erroneous strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05 (Ioannidis, 2005).

Psychological research have relied on statistical incorrect practices (e.g. Null Hypothesis Significance Testing (NHST) approach, and conventional p-value). In addition to the mentioned intrinsically statistical methodological issues, are present fraudulent practices related to the publication process known as *Questionable Research Practices* (QRPs). Consequently, psychological literature requires an inspection.

1.1.1 Statistical Inference

A research, in order to be reproducible and replicable, must be based on statistical methods. Yet, not every sociological research is based on statistical inference. Specifically, two are the statistical classes: descriptive and inferential.

Whereas the first one summarizes and graphs the data recorded in a trial, the latter makes inferences about the wide population by captivating data from a sample. It is essential to stress out that the goal of inferential statistics is to draw conclusions from a sample and generalize them to a population (Elst, 2019), while the descriptive statistics aims to represent the data already obtained. The final goal of Inferential Statistical classes is to reach solid and secure results without testing the whole population, but a sample.

Specifically, two are the types of sociological studies, those might be qualitative or quantitative.

Given (2008), stated that the term *Quantitative Research* refers to approaches to empirical inquiry that collects, analyses, and exhibit data in numerical rather than narrative form, while *Qualitative researches*, vice-versa, exhibit data in a narrative rather than numerical form. Hence, not unexpectedly, quantitative research is often viewed as the antithesis of the qualitative type of research (Given, 2008).

Precisely, as reported by Amaratunga et al. (2002), a quantitative approach is based on the rigid academic investigation of numbers that represents opinions and concepts, whereas a research is qualitative when the former concentrates on words and observations to attempt to describe people in natural situations (Amaratunga et al., 2002). Nonetheless, qualitative research is uniquely precious because it allows the use of Statistical Inference to provide a conceptual and computational framework for addressing the scientific question.

Social science researchers recently switched from qualitative to quantitative researchers since inference statistics can be smoothly applied in the latter case. Therefore, quantitative research, which applies fixed procedures and methods, allows a second experimenter to re-conduct and test it again, boosting reproducibility.

Finally, statistics, with the particular emphasis on Inferential Statistics, might be considered to be the core of scientific research.

1.1.2 Reproducibility and Replication

The term *Reproducibility* states the possibility of a researcher to duplicate the results of a prior study by using the same materials manipulated by the original investigator(s) (Goodman et al., 2016). Briefly, reproducibility is obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis.

Therefore, reproducibility ultimate aim is to attempt to confirm the results obtained in the original study by providing the same conclusions for the same research question in further studies. The goal is achieved by computing the same raw data to build the same analysis and to implement the same statistical analysis (Brockett & Mesarović, 1965).

To provide credibility to scientific claims, *Replication* is entailed. The common understanding of replication considers it as a duplication of a study's procedure by recording new data and by observing whether the prior finding recurs. The given definition of replication is intuitive and easy; despite this, Nosek & Errington (2020) argued that it is incomplete as it puts the emphasis solely on the repetition of technical methods, such as on the procedure, protocol, or manipulated and measured events. Therefore, Nosek & Errington (2020) provided an alternative definition for replication, by including the statement that a replication is a study for which any outcome would be considered diagnostic evidence about a prior claim either way. Consider the following: Nosek & Errington (2020) asserted that individual studies examines only a subset of units, outcomes and setting; as they were conducted in a particular climate, in a particular time of the day, in particular point of history, with a particular measurement method, using particular assessments, with a particular sample.

Consequently, it is evident that it is risky to base a theory solely on a single study as the subset of variables is too narrow to advance an attainable statement. Moving from this assumption, the incredible power of replication is that it allows to advance a theory of a broad effect by confronting existing understanding of what a phenomenon is, with new evidence.

Accordingly, testing whether a theory can be functional in a different subset of variables is when the magic happens.I Ironically, when the existing understanding is weak the value of replication is even more essential (Asendorpf et al., 2013). In fact, by using the same procedures to the original study but different raw data (Goodman et al.,

2016) replicability procedures reduces the uncertainty in the universe of possible units, treatments, outcomes, and settings that could be essential to raise the truthfulness of claim made (Nosek & Errington, 2020).

Hence, the fundamental question reproducibility aids to answer is: *of the innumerable variations in units, treatments, out comes, and settings, which one does matter?*

Finally, replication is about identifying the conditions sufficient for assessing prior claims, and a replication test assesses generalizability to the new study's unique conditions; thus a successful replication provides evidences of generalizability across the conditions that inevitably differ from the original study, while an unsuccessful replication indicates in which conditions the theory is not working (Nosek & Errington, 2020). Yet, successful replications can increase the confidence in the claims being made; on the other hand, unsuccessful ones aid to facilitate theoretical innovation to improve or discard the model.

Shortly, reproducibility is a necessary tool to generate generalizability theories across units, treatments, outcomes, and settings facilitates, as a single study that cannot comprehend the overall evaluation of all of the conditions of the claim. In other words, the boundaries of that claim are tested by future replications by considering different units in the tests.

Reproducibility and Replicability are two concepts strictly related to a third key concept: *Generalizability*. Generalizability refers to the possibility to extend the results obtained from a study into another context of population which differs from the original one. On this point, it is vital to note that data reproducibility is compulsory but not sufficient for replicability, and replicability is obligatory but not adequate for generalizability. Thus, the original study necessarily must provide the elements to allow a second researcher to reproduce it. Once the first point is satisfied, unfortunately there is no security for what concerns the result of the reproduction. Specifically, the reproduction may not successfully produce similar findings when compared to the original study. Additionally, a reviewer might decide to replicate the finding with another sample, again the outcome is not certain, the replication could fail.

Even in the positive case when an original study is successfully reproduced and replicated, this does not imply that the finding can be easily generalized to other contexts (e.g. in other populations) (Asendorpf et al., 2013). Therefore, many are the context in which the finding should be tested in order to generalize the results.

To conclude, a reminder: *the result is not the most important element, what is crucial is the advancement of scientific knowledge about the topic of interest.*

1.1.3 Definitions

Replicability has been defined as a particular case of reproducibility (Klein et al., 2014).

Reproducibility is an inclusive term which can be broken into: methods reproducibility, results reproducibility and inferential reproducibility.

Methods reproducibility or solely "reproducibility" means to reach the same conclusions as in the original study, starting from the same data-set and following the same methods and data analysis procedures. This is also called "checking of the analysis."

Results reproducibility or "replicability" is the using a different sample drawn from the same population while following the same methodology of the original authors. Two are the possible cases:

1. Direct, or Exact replication: researcher performs the same procedure, measurement, and analyses;

2. Conceptual replication: different measurements, manipulations or a combination of these two elements are introduced in the procedure;

The distinction among Direct or Exact replication and Conceptual replication is essential.

A core difference is that conceptual replications, introduce new elements in studies' procedures; consequently an unsuccessful conceptual replication might be due to methodological differences between the original and the new study.

Direct or exact replications aim to recreate the same conditions of the original study; in case incompatible results are obtained, it is possible to make meaningful comparisons between the two studies' results (Bertoldo, 2019).

Inferential reproducibility consists of drawing the same conclusions from a reanalysis of an original study or a replication of a study; it is centered on the conclusions of a study. Specifically, 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. This type of reproducibility is not generally considered, however it is possibly the most informative (Bertoldo, 2019).

For a visual representation, refer to Figure 1.1.

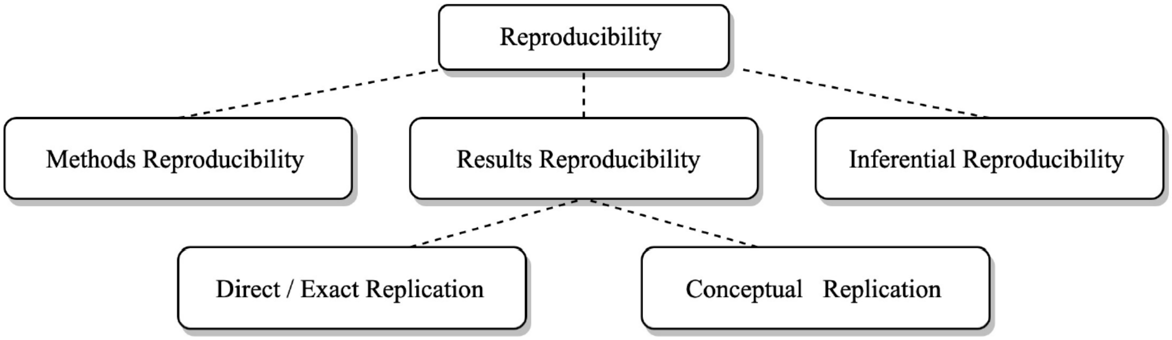


Figure 1.1: Classification of different types of reproducibility by G.Bertoldo, 2019, Dealing with the replication crisis in psychological science: The contribution of Type M and Type S errors

1.1.4 Difficulties when interpreting results

A large replication project suggested psychology is facing a replication crisis. The concern is justified by the existing methodological issues in psychological research. However, conclusions obtained from single replication studies must be careful.

Specifically, a single replication study is likely not sufficient to declare that an effect is nearly equal to zero, reversing the original finding -unless the replication has a surprisingly large sample size-. Therefore, warn about the limitations of single replication studies, chiefly when they sustain the non-existence of a phenomenon (Bertoldo, 2019).

Some of the main problems concern statistical power and the sample size needed to prove that an effect is essentially equal to zero are:

1. How to adequately power a replication study;
2. Sample size needed to reach a certain level of power depends on the size of the effect under study;
3. 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,” whereas *Predictive Power* “acknowledges that the effect size is typically not known with certainty but instead is at best an estimate.” Since predictive power takes account for the uncertainty of the effect size into account, higher sample sizes is recommended.

Sadly, is complicated to be confident to have adequate power, even in replication studies. Therefore, one failed replication with a not large enough sample size, is not evidence for the non-existence of a phenomenon, and multiple studies are needed to prove the absence of phenomenon (Calin-Jageman & Caldwell, 2014).

Some of the solutions proposed to face aforementioned issues are Crowd-Sourcing projects and Registered Replication (Calin-Jageman & Caldwell, 2014).

1.2 Positive effects of the crisis

A positive consequence of the crisis lead to the reward of transparency and an incentive to replicate studies. This is happening through a revision of the statistical method applied in addition, new practices which have been proposed after the crisis, to invert the direction.

In response to the Reproducibility and Replicability issues, scientific studies -including psychological ones- have begun to adopt practices aimed to reduce the frequency of publishing mediocre and inadequate researches. Recently have been created online repositories, such as the Open Science Framework (OSF) where material, data and analyses are uploaded and stored; an unique URL serve to identify them and they are available to scientific community. Therefore, a positive consequence of the increased awareness of Questionable Research Practices (QRPs) is the creation of new practices to reward transparency and incentive replication studies: Pre-Registrations (PRs), Registered Reports (RRs) and Registered Replications Reports (RRR) (Bertoldo, 2019).

- Pre-registration is a tool which allows the authors to upload online their project, where modification of the original project is permitted (e.g. hypothesis, how to test them), however all the changes are tracked and visible in the online accessible file;
- A second from of preregistration are Registered Reports (RRs). The present is an option offered directly by journals where the research process is separated into two stages. In the *first stage*, 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 correct, 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 (Bertoldo, 2019). At this point, there will be a second round of per-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;

- A particular type of article is 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.

1.3 Aim

Last years have been influenced by credibility crisis, since a consistent number of research's findings resulted unclear, unreliable or couldn't be replicated.

In order to minimize sources of errors, the preferential strategy is to increase replicability by decreasing sources of error. Statistical power analysis utilizes the mathematical connection among the four variables (i.e. *power*, α, N, and *ES*) in statistical inference, since the correlation between them allows the expert to determine the fourth element when the other three are set. Asendorpf et al. (2013) suggested that in order to diminish errors, one of the strategies available is to increase Sample size: statistical power goes up and confidence interval (CI) width goes down with larger sample size.

On the other hand, there has been a response to the reproducibility and replicability issues related to publication bias and practices' flexibility, which are examples of *Questionable Research Practices* (QRPs). Specifically, scientific studies -including psychological ones- have begun to adopt practices aimed at enhancing the quality of publication.

On this point recently have been created online repositories, such as the Open Science Framework (OSF) to share data an codes.

Among the three forms of pre-registration, we have decided to deeply analyze Registered Reports (RRs). Specifically, we have taken into account psychological Registered Reports (RRs) pre-registered on Open Science Framework (OSF), as of January 2022 Trough an exploratory study the present paper aims to answer the following question:

Do Registered Reports Improve Sample Size Planning in Psychology?

Registered Reports are recent: they have been proposed in 2012 by Chris Chamber. Therefore, few are the studies which evaluated them. Particularly, to our knowledge, no study focused on the evaluation of Sample Size planning.

Specifically, we are going to:

1. At a descriptive level, evaluate the most common practices to Plan for Sample Size;
2. Individuate points of strength and weakness;
3. Summarize the results obtained to suggest future improvements.

Chapter 2

Power Analysis and Design Analysis

In this chapter, we describe three approaches to statistical inference: the one of Fisher, the one of Neyman and Person, and Null Hypothesis Significance Testing (NHST). We point out that although NHST is an incoherent framework, it has been largely used, therefore it could explain the large amount of statistical misconceptions and the neglect of statistical power and effect size. We give a definition of level of significance, emphasizing that in the present paper we refer to Neyman and Person view of it. Afterwards, we present Statistical Power Analysis from a statistical perspective, additionally we include recommendation to increase power. Additionally, we portray Winner Curse Fallacy: an important heuristic, especially in underpower studies. Finally, we present Design Analysis which allows considering the probabilities of Type M (magnitude) error and Type S (sign) errors in a study.

2.1 Statistical Inference Frameworks

Any research, in order to be reproducible and replicable should be founded on solid scientific knowledge. The application of Statistical Inference is widely used to analyze empirical studies' results.

In the present section are presented three Statistical Inference Approaches: Fisher, Neyman-Pearson and Null Hypothesis Significance Testing (NHST). In regards of Null Hypothesis Significance Testing, Gigerenzer et al. (2004) stated that NHST is an incoherent framework which was born from a mixture of Fisher's Null Hypothesis Testing and Neyman and Pearson Decision Theory.

Despite this, editors of major journals made null hypothesis testing a necessary condition for the acceptance of papers and made small *p*-values the hallmark of excellent experimentation (Gigerenzer et al., 2004); this might partially explain some statistical misconceptions and the neglect of notions such as statistical power and effect sizes.

It is essential to note that the approaches taken into account in the present paper are included in Frequentist statistics; Bayesian statistics although relevant and progressively used, won't be taken into account in this work.

2.1.1 Fisher Approach

Fisher Significance Testing or Null Hypothesis Testing:

- Set up a statistical null hypothesis. The null can't be a nil hypothesis (zero difference);
- Report the exact level of significance (e.g., $p = 0.049$ or p=.051). Do not use a conventional 5% level (e.g., p <.05), and do not talk about accepting or rejecting hypothesis (Gigerenzer, 2004);
- Use this procedure only if you know very little about the problem at hand (Gigerenzer, 2004).

Fisher proposed a method of statistical inference called *Null Hypothesis Testing* or *Significance Testing*.

In the framework, the researcher specifies a hypothesis named *Null*, not necessarily because it postulates an effect equal to zero, which is instead a *Nil Hypothesis* (Bertoldo, 2019); the null hypothesis refers to the hypothesis the researcher aims to nullify. The observations gathered in a study are compared to observations anticipated in the case the null hypothesis is the real state of the world; the final target is to estimate the deviation of the observation from the predicted *Null Hypothesis*. It is feasible to reject the *Null Hypothesis* as a plausible representation of the world when data are not plausible under the *Null Hypothesis* (Gigerenzer et al., 2004).

Fisher introduced the concepts of *significance level* and p-value to indicate how extreme data should be to reject the null hypothesis (Gigerenzer, 2004). Indeed, Fisher proposed an epistemic interpretation of probability supporting that the level of significance of a result could give information about the degree of belief in the null hypothesis (Gigerenzer, 2004). Accordingly, 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, 2004).

Fisher suggested that *Null Hypothesis Testing* is particularly adequate when little or nothing is known about the phenomena of interest and this procedure is not applicable on every scientific research. Although, this situation is rare in psychological science, where usually there is a wealth of competing theories and published studies (Bertoldo, 2019).

2.1.2 Neyman and Pearson Approach

Neyman-Pearson decision theory:

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 mean that you believe in it, but only that you act as if it were true;

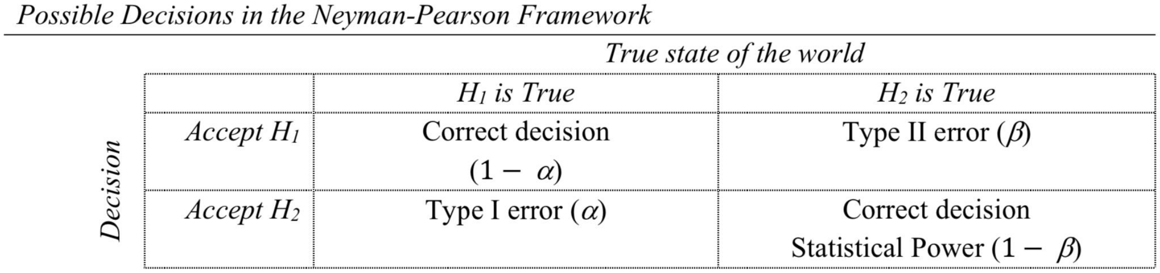
3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g., either my_1 =8 or my_2= 10 is true) and where you can make meaningful cost-benefit trade-offs for choosing α and β.

Neyman and Pearson framework can be seen as a decision theory where the researcher sets two competing hypotheses (H_1 and H_2); when H_1 is called the *Null Hypothesis*, while H_2 is named *Alternative Hypothesis*. Subsequently, it is tested H_1 assuming it is the real state of the world, then H_2 assuming it is the real state of the world: therefore the reference class is used to compute the sampling distribution of the test.

The *threshold* depends on the probability that the experimenter is inclined to accept committing two possible types of errors when repeating this experiment a number of times that tends to infinite:

- To accept H_1 when H_2 is the real state of the world, called Type I error. The probability of committing a Type I error is indicated with α. For instance: when α is set at 5% means 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 be infinite (Bertoldo, 2019).
- To accept H_2 when H_1 is the real state of the world, called Type II error. The probability of committing a Type II error is indicated with β and it relies on the size of the effect under study. For example, when the researcher sets β= 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 be infinite.

Neyman and Pearson approach allows the researchers to takes decision. Every time a decision is taken, it could be right or incorrect; it is therefore vital to maintain the frequency of wrong decision under a certain level: to accomplish this task α and β need to be established (Bertoldo, 2019), as shown in Figure 2.1.



Possible Decisions in the Neyman-Pearson Framework

		True state of the world	
		H_1 is True	*H_2 is True*
Decision	*Accept H_1*	Correct decision $(1 - \alpha)$	Type II error (β)
	Accept H_2	Type I error (α)	Correct decision Statistical Power $(1 - \beta)$

Figure 2.1: Possible decisions in Neyman-Pearson Framework

2.1.3 Fisher and Neyman-Pearson: a comparison

Fisher's null hypothesis testing, at each step is unlike Neyman-Pearson decision theory. It lacks a specified statistical alternative hypothesis. As a consequence, the concepts of statistical power,Type-II error rates, and theoretical effect sizes have no place in Fisher's framework since there is no need to specify any alternative for those concepts. Neyman and Pearson criticized Fisher's null hypothesis testing for several reasons, including that no alternative hypothesis is specified (Gigerenzer, 2004).

Fisher held an epistemic interpretation of probability, where the level of significance of a result could give information about the degree of belief in the null hypothesis. 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 (Bertoldo, 2019).

In its simplest version, Neyman–Pearson theory has two hypotheses and a binary decision criterion (Gigerenzer, 2004). Fisher considered Neyman-Pearson approach inadequate for scientific problems where usually it is not even possible to repeatedly perform random sampling from a precise population, while Neyman and Pearson considered Fisher's approach useless in those cases where some tests had less power than the α level (Bertoldo, 2019).

Anyhow, both parties agreed inference should not be an automatic process, it should be based on informed judgments. Unfortunately, NHST, the hybrid version of the two approaches that became popular among social scientists, promoted this is as a general attitude.

2.1.4 NHST Approach

Null Hypothesis Significance Testing (NHST) was born in the 1940s when the Statistical Inference Revolution happened. Soon it became popular and consequently largely used. As a result, textbooks identified NHST as the *the method* for scientific inference and journal policies began to use NHST to both validate and evaluate research.

NHST is a statistical inference approach in which an exploratory hypothesis is tested against a hypothesis of no effect or no relationship based on a given observation (Gigerenzer et al., 2004). NHST was born by picking key elements of Fisher's approach and blended them with essential characteristic of Neyman-Pearson's framework (Gigerenzer et al., 2004). However, many experts claimed the unreliability of NHST as it is the illogical combination of Fisher's framework with Neyman-Pearson's one, which cannot not be integrated (Gigerenzer, 2004).

Gigerenzer et al. (2004) summarized the mechanical procedure used in NHST is labeled *Null Ritual* which consists in:

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

A consequence of this inattentive and automatic procedure, results are easily biased and researchers might misinterpret the findings in two possible ways (Gigerenzer et al., 2004):

- firstly by going beyond the "rejection of H_0" such as claiming for a statistically significant result (i.e. p-value lower than the critical value) that the alternative hypothesis is true;
- or they can "fail to reject H_0" by holding for non-statistically significant result (i.e. 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*).

NHST is characterized by statistical misconceptions and the neglect of notions such as statistical power and effect sizes which lead to impaired consequences for the robustness of findings. In fact, only the null hypothesis is formalized, whereas the alternative hypothesis is not specified in statistical terms. The consequent problem which arises is that the null hypothesis is likely to become a *straw man* when its rejection is taken as a way to accept the alternative hypothesis (Loken & Gelman, 2017). In reality, when an alternative hypothesis is introduce and tested, Type II error should be controlled and statistical power should be considered. The lack of reflection on power likely leads to underpowered studies and absence of awareness of the associated risks.

Specifically, underpowered studies are researches where the probability of being able to reject H_0 is very low, if H_0 is actually false; such as even when the effect of interest exists, the study has a low probability to find it.

Over the last decade, researchers have focused on reproducing and replicating published experiments in order to validate them, however, the majority of studies did not reproduce and neither replicate and this led to a crisis in psychology; partially as a result of mindless application of NHST.

Hence, results from underpowered studies should be carefully scrutinized (Loken & Gelman, 2017).

2.1.5 Level of Significance

According to Gigerenzer et al. (2004), *level of significance* has three meanings:

- the conventional level of significance, a common standard for all researchers (early Fisher). You specify only one statistical hypothesis, the null. You always use the 5% level and report whether the result is significant or not; that is, you report p $< .05$ or $p > .05$, just like in the null ritual. If the result is significant, you reject the null; otherwise, you do not draw any conclusion. There is no way to confirm the null hypothesis. The decision is asymmetric (Gigerenzer et al., 2004);
- the α level that is, the relative frequency of wrongly rejecting a hypothesis in the long run if it is true, to be decided jointly with β and the sample size before the experiment and independently of the data (Neyman and Pearson). You specify two statistical hypotheses, H1 and H2, to be able to calculate the desired balance between α, β and the sample size N. If the result is significant (i.e., if it falls within the alpha region), the decision is to reject H_1 and to act as if H_2 were true; otherwise, the decision is to reject H_2 and to act as if H1 were true. (We ignore here, for simplicity, the option of a region of indecision.) For instance, if $\alpha = \beta =$ .10, then it does not matter whether the exact level of significance is .06 or .001. The level of significance has no influence on α. Unlike in null hypothesis testing with a conventional level, the decision is symmetric (Gigerenzer et al., 2004);
- the exact level of significance, calculated from the data after the experiment (late Fisher). You calculate the exact level of significance from the data. You report, say, $p = .051$ or $p = .048$. You do not use statements of the type “$p < .05$” but report the exact (or rounded) value. There is no decision involved. You communicate information; you do not make yes-no decisions (Gigerenzer et al., 2004).

The basic difference is this: For Fisher, the exact level of significance is a property of the data, that is, a relation between a body of data and a theory; for Neyman and Pearson, α is a property of the test, not of the data. Level of significance and α are not the same thing.

In this thesis, when not differently specified, we will refer to Neyman and Pearson approach. We chose this approach because in N-P framework is included an Alternative Hypothesis which must be formalized by the researcher. Moreover, power assessment and consequent sample size planning is explicit.

2.2 Statistical Power

In order to minimize sources of errors, the preferential strategy is to increase replicability by decreasing sources of error. Scientists ideally would like to make no errors of inference, that is, they would like to infer from a study a result that is true in the population. If the result is true in the population, a well-powered replication attempt (as discussed later) will likely confirm it (Asendorpf et al., 2013).

Neyman and Pearson approach is the widely used for study design and data analysis, since it is the framework takes into account an alternative hypothesis which have to be formalized by the researcher.

1. rejecting the null hypothesis when it is true (false positive, α);
2. and failing to reject it when it is false (false negative, β).

These two types of errors can be best understood from the perspective of power (Cohen, 1988).

The power of a statistical test is the probability that H_0 will be rejected when it is false, thus power is the probability of obtaining a statistically significant result (Gigerenzer et al., 2004).

Statistical Power depends on three elements:

- Significance criterion (α or β; such as Type I or Type II errors respectively);
- Sample Size (n);
- Population effect size (ES).

Statistical power analysis utilizes the mathematical connection among the four quantities (i.e. *power*, α, n, and ES) in statistical inference, since the relationships between them allows the expert to determine the fourth element when the other three are set (Gigerenzer et al., 2004).

It is possible to minimize both types of errors simultaneously to increase statistical power. Because replicable result are more likely when power is high, the core process is identifying the factors that increase statistical power; that is, for any α level, statistical power increases as effect size and sample size increase.

2.2.1 Recommendations

In order to increase statistical power, Asendorpf et al. (2013) suggested that it is useful to:

- Increase Sample size: statistical power goes up and confidence interval (CI) width goes down with larger sample size. Therefore, results obtained with larger samples are more likely to be replicable than those obtained with smaller ones. This has been said many times over the decades, however reviews have shown little improvement in the typical sample sizes used in psychological studies. Publishing many low-powered studies contributes to this excessive false-positive bias. It cannot be stressed enough that researchers should collect bigger sample sizes, and editors, reviewers, and readers should insist on them.
- Increase study design sensitivity, such as having better control over methodological error sources: This means distinguishing between systematic and random errors. While the first error type is challenging to act on as they have no explanation and it is appalling to predict them, the second error type can be reduced by using clear and standardized instructions, paying attention to questionnaire administration conditions, and using stronger manipulations in experimental designs.

An additional element Asendorpf et al. (2013) focused on was to increase adequacy of statistical analyses, thus choosing an appropriate statistical analysis that fits the design can decrease errors. Lastly, errors can be reduced by establishing whether finding is replicated and by defining whether the core parameters are statistically significant when compared to the origin replicated study.

- Increase reliability of the measures: The two most common estimators of effect size (Cohen's *d* and Pearson's *r*) both have standard deviations in their denominators; hence, all else equal, effect sizes go up and CIs and standard errors down with decreasing standard deviations. Because standard deviation is the square root of variance, the question becomes how can measure variance be reduced without restricting true variation? The answer is that measure variance that can be attributed to error should be reduced. This can be accomplished by increasing measure reliability, which is defined as the proportion of measure variation attributable to true variation. All else equal, more reliable measures have less measurement error and thus increase replicability;
- Increase adequacy of statistical analyses: testing appropriateness of method-required assumptions, treating stimuli as random rather than fixed factors (Judd et al., 2012), respecting dependence within the data, and removing the influences of covariates, given appropriate theoretical rationale;
- Avoid multiple underpowered studies: a series of underpowered studies with the same result are so unlikely that the whole pattern of results becomes literally 'incredible.' It suggests the existence of unreported studies showing no effect. Even more, however, it suggests sampling and design biases. Such problems are very common in many recently published studies;
- Consider error introduced by multiple testing: the likelihood that some among multiple variables will show significant relations with another variable is higher with underpowered studies, although the likelihood that any specific variable will show a significant relation with another specific variable is smaller. Consequently,

> the literature is scattered with inconsistent results because underpowered studies produce different sets of significant (or nonsignificant) relations between variables.

Even worse, it is polluted by single studies reporting overestimated effect sizes, a problem aggravated by the confirmation bias in publication and a tendency to reframe studies post hoc to feature whatever results came out significant. Statistical, as well as non-statistical solutions are available, the latter propose a explicit separation of a priori hypotheses preregistered in a repository from exploratory post hoc hypotheses.

2.2.2 Planning for Sample Sizes

The core goal of sample size justification for empirical studies is to explain how the collected data is expected to provide valuable information given the inferential goals of the researcher (Lakens, 2021); therefore, sample size justification is an essential step while designing those studies.

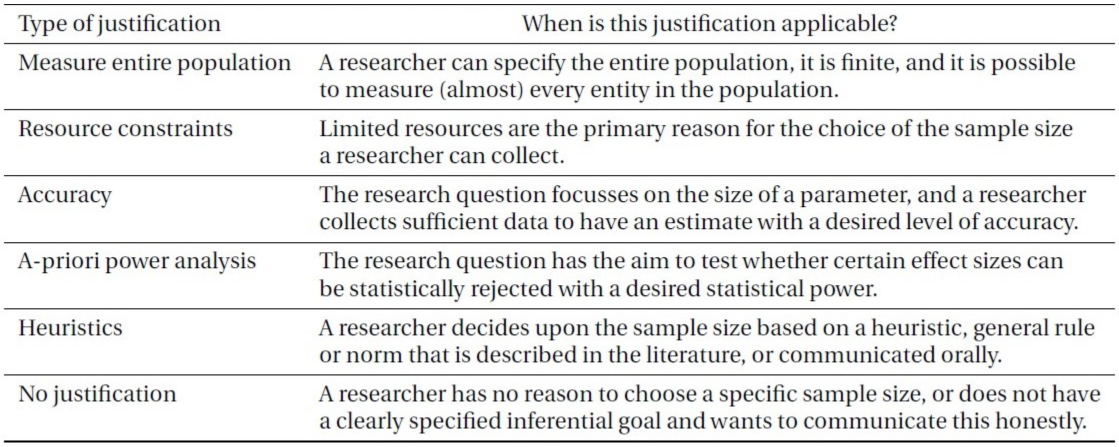
Turner et al. (2013) have studied the impact of study size on meta-analyses. Specifically, most meta-analyses include data from small studies that individually do not possess the power to detect an intervention effect: 14886 meta-analyses have been review, and in 10492 cases -which amount corresponds to the 70%- they were underpowered (Turner et al., 2013).

Furthermore, has been shown that the median sample size reported in papers published in representative journals are around n=40 and the effect size found in the meta-analyses are around d=0.50 (Asendorpf et al., 2013).

On those points, Bakker et al. (2020a) recently published an article named *Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size*, where has been reported an investigation aimed to assess whether the statistical power of a study is higher when researchers are asked to make a formal power analysis before collecting data.

Lakens (2021) offered an overview of the most diffuse approaches to justify sample size for single studies.

For a visual representation, refer to Figure 2.2.



Type of justification	When is this justification applicable?
Measure entire population	A researcher can specify the entire population, it is finite, and it is possible to measure (almost) every entity in the population.
Resource constraints	Limited resources are the primary reason for the choice of the sample size a researcher can collect.
Accuracy	The research question focusses on the size of a parameter, and a researcher collects sufficient data to have an estimate with a desired level of accuracy.
A-priori power analysis	The research question has the aim to test whether certain effect sizes can be statistically rejected with a desired statistical power.
Heuristics	A researcher decides upon the sample size based on a heuristic, general rule or norm that is described in the literature, or communicated orally.
No justification	A researcher has no reason to choose a specific sample size, or does not have a clearly specified inferential goal and wants to communicate this honestly.

Figure 2.2: Overview of possible justifications for the sample size in a study. Sample Size Justification by D.Lakens et al., 2019

- The first justification includes data from (nearly) the whole population of interest: census;

In some conditions, researchers are able to collect data from (almost) the totality of the population of interest. For example: a study wants to analyze data from a small group of people, for instance Italian elderly people of 115 years old.

When it is possible to measure the entire population, the sample size justification becomes straightforward: the researcher used all the available data. Moreover, when the entire population is measured absent is the need to perform a hypothesis test: the generalization is superfluous, there is no population to generalize to, everyone has been directly tested.

- Moving forward, the second justification found itself on resource contains, which are almost inevitably always present, however rarely are explicitly evaluated;

Despite the omnipresence of resource limitations, the topic often receives little attention in texts on experimental design. Resource constraint justifications are based on a trade off between the costs of data collection, and the value of having access to the information the data provides. Two of the constraints all the scientists face are time and money (Lakens, 2021).

- Both the third and fourth justifications are based either on a desired statistical power or a desired accuracy;
- A-priori Power Analysis;

When designing a study where the goal is to test whether a statistically significant effect is present, researchers often want to make sure their sample size is large enough to prevent erroneous conclusions for a range of effect sizes they care about. In this approach to justifying a sample size, the value of information is to collect observations up to the point that the probability of an erroneous inference is, in the long run, not larger than a desired value (Lakens, 2021).

When a researcher performs a hypothesis test, there are four possible outcomes:

1. A false positive (or Type I error);
2. A false negative (or Type II error);
3. A true negative;
4. A true positive.

Given a specified effect size, alpha level, and power, an a-priori power analysis could be used to calculate the number of observations required to achieve the desired error rates, given effect size (Lakens, 2021).

- Planning for Precision;

Some researchers have suggested to justify sample sizes based on a desired level of precision of the estimate. The goal when justifying a sample size based on precision is to collect data to achieve a desired width of the confidence interval around a parameter estimate. The width of the confidence interval around the parameter estimate depends on the standard deviation and the number of observations. The only aspect a researcher needs to justify for a sample size justification based on accuracy is the desired width of the confidence interval with respect to their inferential goal, and their assumption about the population standard deviation of the measure (Lakens, 2021).

- The fifth justification relies on heuristic;

When a researcher uses a heuristic, authors are not able to justify their sample size themselves. Contrary, experimenters trust in a sample size recommended by some authority. These rules of thumb seem to primarily emerge due to mis-citations and/or overly simplistic recommendations. A further popular heuristic is to collect the same number of observations as were collected in a previous study. Using the same sample size as a previous study is only a valid approach if the sample size justification in the previous study also applies to the current study. The suggestion is that instead of stating that you intend to collect the same sample size as an earlier study, one should repeat the sample size justification, and update it accordingly to any new information (Lakens, 2021).

- Lastly, researchers can select a sample size without any justification.

The last category is the one where the author explicitly states they do not have any justification for their sample size. The absence of sample size justification is not necessarily bad. It is still possible to discuss about the data collected (e.g. the smallest effect size of interest, the minimal statistically detectable effect, the width of the confidence interval around the effect size, and to plot a sensitivity power analysis) in relation to the sample size that was collected. If a researcher truly had no specific inferential goals when collecting the data, such an evaluation can perhaps be performed based on reasonable inferential goals peers would have when they learn about the existence of the collected data (Lakens, 2021).

2.2.3 Effect size

Effect Size is an indicator of the magnitude of an effect. Precisely, Cohen (1988) defined effect size as the *degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false.* A plausible effect size is an estimation of the results of an experiment before conducting it which is provided by the researchers themselves: estimations of the effect size are informative for what concerns the practical and theoretical importance of an effect (Fritz et al., 2012).

The two most commonly used measures of effect size are Cohen's d and Pearson's r. The former, typically used to characterize the differences in means between independent experimental groups, is the mean difference divided by the pooled standard deviation. The latter, the correlation coefficient, is typically used to characterize the degree to which one variable can be predicted from another (Funder & Ozer, 2019) In Cohen (1988) words, specifically, the first family is useful to express group differences, while the second to express association between variables.

Cohen (1988) interpretation of effect size is based on thresholds, which provided criteria to understand the effect sizes calculated by suggesting indicative values of d for *large*, *medium*, and *small* effect sizes.

- A large effect size corresponds to d=0.8 and it refers to very obvious differences;
- A medium effect size corresponds to d=0.5 and it refers to differences that are *large enough to be visible to the naked eye* (Cohen, 1988);
- Finally, a small effect size corresponds to d=0.2 and it refers to small differences that are difficult to detect;

However, The terms small, medium, and large are meaningless in the absence of a frame of reference (Funder & Ozer, 2019).

Despite the core value of predicting a plausible effect size, Fritz et al. (2012) pointed out that effect sizes were reported for fewer than half of the analyses in articles published in 2009 and in 2010 in the *Journal of Experimental Psychology: General*, which is a key source of knowledge for psychologists all over the world. Even when effect sizes are reported, are too often underappreciated and misinterpreted. The most common mistakes are describing them in ways that are uninformative (e.g., using arbitrary standards) or misleading (e.g., squaring effect-size) (Funder & Ozer, 2019).

Funder & Ozer (2019) proposed that effect sizes can be usefully evaluated by comparing them with well-understood benchmarks or by considering them in terms of concrete consequences. Further, all estimations of effect size should be evaluated in the context of the research. It is not sensible to say of some phenomenon that its effect size is X without specifying under what conditions it has been found to be X (Fritz et al., 2012).

Funder & Ozer (2019) assessed how can effect sizes be interpreted in a way that adds or provides meaning, through:

1. Benchmarks: The idea behind using benchmarks to evaluate effect size is that the magnitude of a finding can be illuminated by comparing it with some other finding that is already well understood;

2. Consequences: A more direct way to evaluate an effect size is to consider consequences, which in some cases can be numerically calculated. Perhaps the best known and easiest to use of these methods is the bi-nominal effect-size display (BESD).

Furthermore, Funder & Ozer (2019) evaluate effect size implications for how research findings should be interpreted:

1. Researchers should not automatically disregard *small* effects: social psychologists have seemed reluctant to report or to emphasize effect sizes. P-hacking is a structure which incentives structure that rewards performing selective analyses in order to increase small effect sizes so they cross the threshold of statistical significance. A replace could be a structure which instead incentives gathering data from large

samples and merciless reporting small, effect sizes that are precise and reliable (Funder & Ozer, 2019);

2. Researchers should be more skeptical about *large* effects: researchers have often reported anomalously large effect sizes in small N studies. Because the confidence intervals (CI) of effect sizes in small studies are very wide, such studies can be expected to sometimes produce large apparent effects that replication studies reveal to be greatly overestimated (Funder & Ozer, 2019);
3. Researchers should be more realistic about the aim of their programs of psychological research: it is not realistic to expect that any one research program, on any one topic or psychological process, determines more than a small piece of what is really going on in the psychological world (Funder & Ozer, 2019).

Finally, is important to note that, whatever the procedures, all assumptions that will lead to the identification of a plausible effect size must be communicated in a transparent manner; such as by increasing the information provided by a study and ensuring more reasonable statistical claims related to the obtained results, whether they are significant or not (Altoè et al., 2020).

2.2.4 Justify your Alpha

It has been seen previously that Fisher introduced the concepts of *significance level* and p-values: indicators of the extremeness required for the data to reject the *Null Hypothesis*.

Firstly Fisher established the level of significance at 5% or, alternatively 1%. That is, when the probability of the data observed and collected under the *Null Hypothesis* was inferior of 5%, that is p-value < 0.05, then result could be defined as *statistically significant*, consequently it leads to the rejection of the *Null Hypothesis* (Bertoldo, 2019).

Later, Fisher himself, stated that 5% level of significance should not be used in all the contexts, and scientists should report the exact *p*-value obtained in a study.

Similarly to the last Fisher statement, decades later Lakens et al. (2018) proposed that researchers should transparently report and justify choices made while designing a study, *before running the experiment*, including alpha level.

Specifically, Lakens et al. (2018) strongly asserted that the label *statistically significant* should no longer be used. In contrast, researchers should provide a meaningful interpretation of the theoretical or practical relevance of their results, without relying on a priori determined threshold. Moreover, according to Lakens et al. (2018), while designing a study, authors should transparently specify and justify their choices.

Finally, providing researchers (and reviewers) with accessible information about ways to justify (and evaluate) design choices, tailored to specific research areas, will improve current research practices.

2.2.5 Is time spent on testing hypotheses worthy?

Scheel et al. (2021), recently proposed that prior to hypothesis testing, experimenters ought dedicate effort in forming concepts, developing valid measures, establishing the causal relationships between concepts and the functional form of those relationships, additionally to identifying boundary conditions and auxiliary assumptions. Therefore, every researcher who can provide these inputs should be boosted since they are significantly aiding in the progress of science. The shift of focus advanced by Scheel et al. (2021) is in sharp contrast with the willingness to test hypotheses, that in the 1940s exacerbated in the mindless use of NHST. As a consequence of the replication crisis, psychological reforms have narrowed their attention on formalizing procedures from testing hypotheses. Scheel et al. (2021) stated that although these reforms were necessary, psychologists have realized they might not be ready to test hypotheses. Finally, the authors concluded that for the scientists in psychological field, the premature test of hypothesis before establishing a *derivation chain* between test and theory could be easily pointless and disadvantageous (Scheel et al., 2021).

2.2.6 Winner Curse

Winner Curse Fallacy is an important heuristic, especially in underpower studies.

A consequential common misinterpretation of the results obtained by low power studies is a bias named by Loken & Gelman (2017) *what does not kill statistical significance makes it stronger fallacy*, in which analyzers might consider the result even more remarkable when a statistically significant result is found despite the low probability of this to happen (e.g., low power).

In particular, it is considered a fallacy since it is possible to accomplish statistical significance because of many other factors that may be involved in the sampling in addition to the effect the study is interested in. Some of the factors are: researchers degrees of freedom, large measurement errors and **small sample sizes** (Altoè et al., 2020).

Thereupon, the apparent win in terms of obtaining a statistically significant result is actually a loss, since the *the lucky* scientist who made a discovery is *cursed* by finding an inflated estimate of that effect. For the above mentioned reason, the fallacy is known as *Winner Curse* (Button et al., 2013).

In brief, the lack of statistical power might lead to misinterpretation of the results, in addition in addition to share overestimate of effect size in the literature.

2.3 Design Analysis

The world of psychological research is fascinating but yet impervious. Researches may face inferential risks while evaluating the results of underpowered studies based on statistical significance thresholds. In this scenario, experimenters could find: an exaggerated estimate of the effect, a statistically significant result in the wrong direction, or both.

Statistical issues, in addition to publication biases, are a deadly combination which might explain why the majority of literature effects are biased upwards and replication studies mainly find smaller effect sizes.

Underpowered studies are a feasible explanation for psychology's credibility crisis. In addition, there is the diffuse tendency of researchers to rely blindly on statistical significance as a thresholds to make decision which might be another explanation for the crisis. Either of them, or a combination of those, might conduct to misleading results: the risk of finding an exaggerated estimates of the effect size increase in an underpower study.

Bertoldo (2019) pointed out that apart from the traditional Type I and Type II errors, there are two additional errors to consider. Specifically, in 2014 Gelman and Carlin (Gelman & Carlin, 2014) formalized Type M and Type S errors considered respectively the magnitude and sign of effect sizes.

On one hand, *Type M* (magnitude) error, or *Exaggeration Ratio*, shows the predictable average overestimation of an effect that emerges as statistically significant. On the other hand, *Type S* (sign) error represents the probability that a statistically significant result has the opposite sign of the plausible true effect size.

The analysis called *Design Analysis* takes into account both design studies (Altoè et al., 2020), in addition to estimations of uncertainties with the aim to propose more sensible and logical statistical claims (Altoè et al., 2020).

Ideally, design analysis are performed *before* running the study, as a simple size planning strategy: *prospective design analysis.* Nonetheless, design analysis works efficiently even *after* a study has been run to evaluate it: *retrospective design analysis* (Altoè et al., 2020).

On the top of that, design analysis identifies the plausible magnitude and direction of the effect under study: a *plausible effect size.* Both errors are indeed defined starting from its formalization (Bertoldo et al., 2020).

2.3.1 Type M error, Type S error and Statistical Power

Type M (Magnitude) error and Type S (Sign) are two index formalized by Gelman & Carlin (2014), whose attempt was to quantify the risk to find an exaggerated estimates of the effect, especially in underpower studies which are more vulnerable.

Type M error could be defined as the predictable average overestimation of an effect that emerges as statistically significant. It is vital to note that this inferential risk is not a probability but rather a ratio between the absolute value of the mean of all the effect sizes that are statistically significant and the plausible true effect size. Moreover, it indicates the average percentage of inflation (Bertoldo et al., 2020). Consequently, if there is no exaggeration of the effect that emerges as statistically significant, the value that is obtained is one (i.e., minimum obtainable value). Ideally, Type M error should be as close to one as possible to increase the probability of finding an accurate estimate of the effect.

Type S error is the probability of finding a statistically significant result in the opposite direction to the plausible one. It is calculated by making the ratio between the probability to find a negative statistically effect size and the power of a test, when the plausible effect size is positive. Otherwise, the ratio is between the power of a test and the probability to find a positive statistically significant effect (Bertoldo et al., 2020).

Type M and Type S errors are intrinsically connected to power as shown in Figure 2.3: the highest the power, the lower is the probability to commit a Type S or a Type M errors, as shown in Figure 2.3.

Interestingly, Type S increases exponentially when the power is 20% or below, whereas Type M error increases exponentially when the power goes below 50%. In

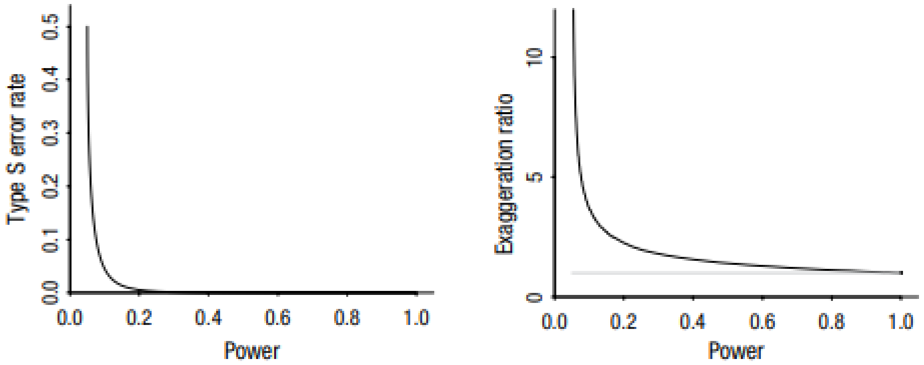


Figure 2.3: Relationship between power, Type S error and Type M error. Reprinted from Beyond Power Calculation: Assessing Type S and Type M Errors, by Gelman, A. et al, 2004, Perspectives on Psychological Science, 9, 6, 644.

the Figure 2.3 it is feasible to appreciate that high level of power required to minimize Type M and Type S errors.

Gelman & Carlin (2014) noted that in any experimental design, statistical power depends on sample size, measurement variance, the number of comparisons being performed, and the size of the effects being studied; although generally sample size is the factor that receive more attention.

Despite usually researchers focus on Type I and Type II errors, evaluation of Type M and Type S errors might increase researchers' comprehension of inferential process by illustrating risks related to Sample Size planning. A study cannot be evaluated solely on its significance; it is required an assessment of its expected power based on a reasonable effect size.

Thanks to Design Analysis, it is possible to estimate an exaggerated estimate of the effect, a statistically significant result in the wrong direction, or both. This estimation is possible through the calculation of the probabilities of committing a Type M (magnitude) error, a Type S (sign) error, plus traditional Type I and Type II errors in a study.

Chapter 3

Registered Reports

In this chapter, we describe Questionable Research Practices (QRPs): methodological and statistical practices that bias the scientific literature and affect credibility and reproducibility of research findings. with a focal point on P-Hacking and HARKing, two of the most common Questionable Research Practices. We analyze a specific form of pre-registration called Registered Reports (RRs), a protocol introduced in 2012: specifically, It has been observed that pre-specifying hypotheses and analyses strategies before examining data which scientific studies -including psychological ones- might reduce the likelihood to publish inadequate researches, including P-Hacking and HARKing. We present Open Source Framework (OSF) as one of the platforms available to pre-register Registered Reports, that we used to gathered that for our dataset. To conclude the chapter, we assess the value of Registered Reports by taking into account the findings of the first studies conducted on this topic.

3.1 Questionable Research Practices (QRPs)

Questionable Research Practices (QRPs) are methodological and statistical practices that bias the scientific literature and affect credibility and reproducibility of research findings.

A survey in the United States revealed that an alarmingly large percentage of university psychologists admitted having used questionable research practices that can contaminate the research literature with false positive and biased findings. The percentages of Italian psychologists who admitted to have used questionable research practices were similar to the results obtained in the United States. The similarity of results obtained in the United States, Italy, and a related study conducted in Germany suggest that adoption of these practices is an international phenomenon and is likely due to systemic features of the international research and publication processes (Agnoli et al., 2017).

Results from replication studies have been considered a red flag for psychology research methods, an additional contemplation on published literature gave rise to the awareness of issues, also knows as credibility crisis.

As Bertoldo (2019) pointed out, it is hard to believe that in a sample of studies published in 1986 and 1987, 96% of results where associated with a p-value lower than 5%, considering statistical power in psychology is low.

A possible explanation is the presence of *publication bias* that favors *positive* over *null* findings; additionally, low p-value can be achieved by exploiting *researchers' degrees of freedom.*

Publication bias and Practices' flexibility are examples of *Questionable Research Practices* (QRPs) which contribute to create findings that upon replication may not hold up.

On one hand, flexibility allows data exploration, which is the core of science as it can lead to new ideas and discoveries (Nosek & Lakens, 2014); all researchers should fully examine and explore their data as it often helps to uncover unexpected patterns that might be the object of further study (Lindsay et al., 2016).

On the other hand, when a practice is too flexible, researchers are able to adjust their measures or designs while analyzing their data and then selectively report just those outcomes that best support the hypothesis. The flexibility to choose which statistical tests to conduct after inspecting the data, dramatically increases the chances of erroneously rejection of Null hypotheses (Type I errors); therefore it's easy to obtain statistical significance when in reality, there is no effect at all.

To summarize: while flexibility might aid data exploration, it also simplifies the cheating game on the result obtained.

About flexibility, Lakens (2019) symbolically reported: *One of its numerous processes is to make multitudes of observations, and out of these to select those only which agree or very nearly agree. If a hundred observations are made, the cook must be very unlucky if he can not pick out fifteen or twenty that will do up for serving.*

QRPs take advantage of flexibility to inflate the false positive rate of published results. As a consequence, QRPs lead to exaggerate the estimations of the size of real effects, which increases the likelihood that subsequent replication attempts will be underpowered and at high risk of failure, and meta-analytic summaries of the literature are less accurate (Lindsay et al., 2016).

It should come clear that proper statistical inference requires full reporting and clarity.

3.1.1 P-Hacking and HARKing

Nowadays, the current publication system is intrinsically problematic because of two types of QRPs: *P-hacking* and *HARKing*.

The first type is *P-hacking*, also known as *Selective Reporting*.

In the scenario, data are manipulated in order to obtain a significant effect. Specifically, P-hacking occurs when researchers collect and select data or statistical analyses until nonsignificant results become significant. For instance, researchers try out several statistical analyses and/or data eligibility criteria and then selectively report those that produce significant results (Head et al., 2015). Some of the most common practices that lead to P-hacking include: conducting analyses midway through experiments to decide whether to continue collecting data, recording many response variables and deciding which to report post-analysis, and deciding whether to include or drop outliers' post-analyses. (Head et al., 2015).

The second one is *HARKing* is the acronym of: *Hypothesizing After the Results are Known* (Lakens, 2019).

A post hoc hypothesis (i.e., based on, or informed by, one's results) is presented in one's research report as if it were based on a priori hypotheses. Hence, in HARKing the results selected from exploratory analyses often were reported as if they were theory-driven and planned (Lakens, 2019).

Head et al. (2015) asserted that quantifying *P-hacking* and *HARKing* is essential because publication of false positives obstruct scientific progress. In many fields, there is a lack of incentives to replicate research; even in a better scenario when a research is replicated, early positive studies often receive more attention than later negative ones. The logical consequence is that false positives can inspire investment in fruitless research programs, and even discredit entire fields (Head et al., 2015).

For a visual representation, refer to Figure 3.1.

3.2 Recommendations for the publication process

A solution to reduce P-hacking and other forms of hidden flexibility (including HARKing) is to pre-specify hypotheses and analyses strategies before examining data: in response to the reproducibility and replicability issues, scientific studies -including psychological ones- have begun to adopt practices aimed at reducing the frequency of publishing inadequate researches.

One of the most common strategy for this purpose is Pre-Registration.

Preregistration guarantees that the reader can distinguish between strategies inde-

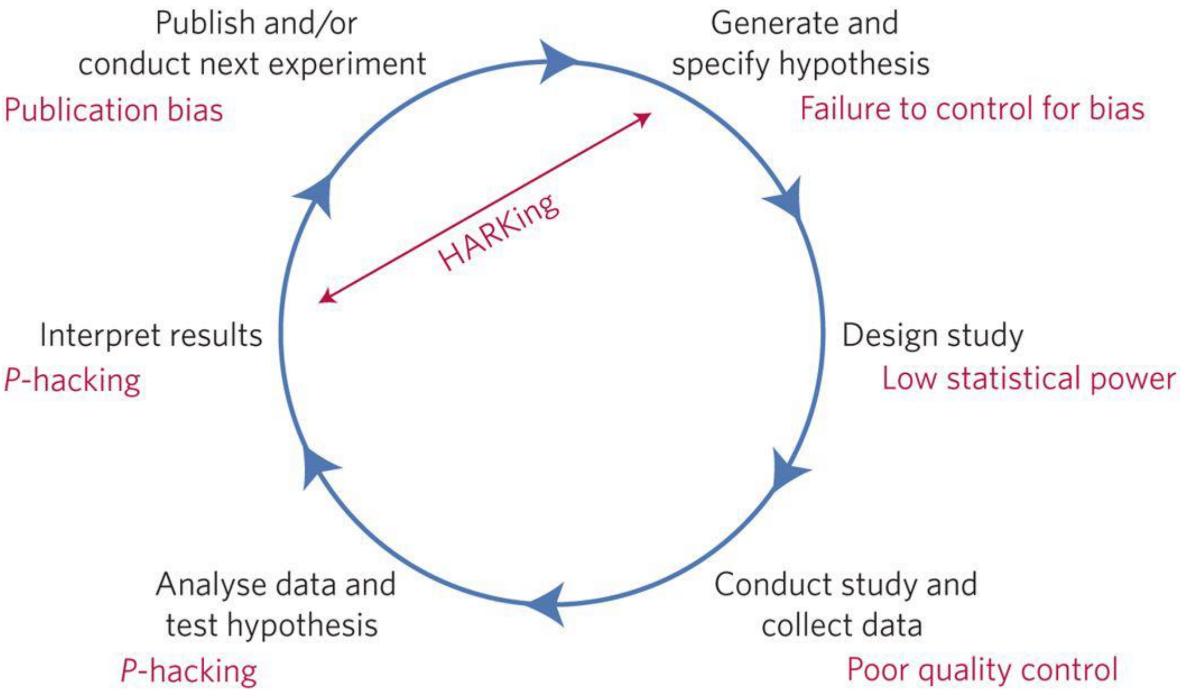

Publish and/or conduct next experiment
Publication bias
Generate and specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and test hypothesis
P-hacking
Conduct study and collect data
Poor quality control

Figure 3.1: 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

pendent from the data, from the ones that were -or might have been- data led. Specifically, data led strategies are not inherently incorrect or misleading, some remarkable advancement in science derived from data exploration, and analytic flexibility. The core issue arises when flexibility is covered up, and the reader is not aware of it. By unmasking this process through preregistration, scientific findings could be more reliable.

Recently have been created online repositories, such as the Open Science Framework (OSF) where material, data and analyses are uploaded and stored; an unique URL serve to identify them and are available to scientific community. Therefore, a positive consequence of the increased awareness of Questionable Research Practices (QRPs) is the creation of new practices to reward transparency and incentive replication studies. Among those practices, we can find: Pre-Registrations (PRs), Registered Reports (RRs) and Registered Replications Reports (RRR) (Bertoldo, 2019).

- Pre-registration is a tool which allows the authors to upload on online repositories their project. Changes of the original project is permitted (e.g. hypothesis, how to test them), however all the changes are tracked and visible in the files which are accessible online;

- A second from of preregistration are Registered Reports (RRs), introduced in 2012. The present is an option offered directly by journals where the research process is separated into two stages. In the *first stage*, researchers submit to the journal the introduction detailing the rationale behind a study, the method section, analysis plan and expected results. A first peer-review process takes place at this stage. If the study is considered theoretically and methodologically correct, 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 (Zgonnikov et al., 2019). 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 on a journal *regardless of the results obtained.* If changes 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.
- A third form of preregistration are 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.

In the present paper we focus on the second form of preregistration aforementioned: Registered Reports.

In order to escalate the limpidity of a study, Asendorpf et al. (2013) recommended that authors should provide: a comprehensive literature review, offer a report sample size decision with a *priori justification for the sample size used*, preregister research predictions and publish materials, data, and analysis scripts.

To accelerate scientific progress authors ought to: publish working papers, conduct replications, engage in scientific debate in online discussion forums.

Reviewers, editors, and journals could: avoid to discourage maintenance of good practices while proactively encourage maintenance of good practices.

Researchers of research methods and statistics might: establish a standard of good practice that values soundness of research over publishability, teach concepts necessary to understand replicable science (e.g. teach and practice rigorous methodology by focusing on multiple experiments), encourage transparency, conduct replication studies in experimental methods classes), critical thinking (e.g. critical reading, critical evaluation of evidence (single-study level) Asendorpf et al. (2013)].

Finally, institutional incentives should focus on quality instead of quantity of publications, use funding decisions to support good research practices, revise tenure standards, and Change informal incentives (Asendorpf et al., 2013).

3.3 Registered Reports

Registered Reports (RRs) have been introduced in 2012 as a way to free researchers from the pressure to engage in these counterproductive practices, thereby breaking the cycle that perpetuates bias and absence of reproducibility. The RR model originates from the simple philosophy that to defeat the distorting effects of outcome bias on science, we must focus on the process and blind the evaluation of science to research outcomes.

In the Registered Reports publication format preregistered study proposals are reviewed before the data is collected.

Manuscripts are published as long as the approved proposal is followed, regardless of the outcome of the results, which prevents publication bias on the part of the journal (Lakens, 2019). Registered Reports have been adopted by more than 200 journals; underlying that this practice has become increasingly popular in a short amount of time (Lakens, 2019).

RRs are a publication format with a restructured submission time-line. Before collecting data, authors submit a study protocol containing their hypotheses, planned methods, and analyses pipeline, which undergoes peer review. If successful, the journal commits to publish the final article following data collection regardless of whether the hypotheses are supported (in-principle acceptance). The authors then collect and analyze the data and complete the final report. The final report is peer reviewed again but, this time, only to ensure that the registered plan was adherent to and stated conclusions are justified (and, if applicable, that the data pass pre-specified quality checks). RRs thus -reporting Scheel (2021), words "combine an antidote to QRPs (preregistration) with an antidote to publication bias because studies are selected for publication before their results are known."

In other words, Bakker et al. (2020b) defines RRs as a way of implementing a review procedure which is peer-reviewed before collecting the data and are published independently of the final results. In this way, registered reports help to prevent low statistical power, selective reporting of results, and publication bias (Bakker et al., 2020b).

For a visual representation, refer to Figure 3.2.

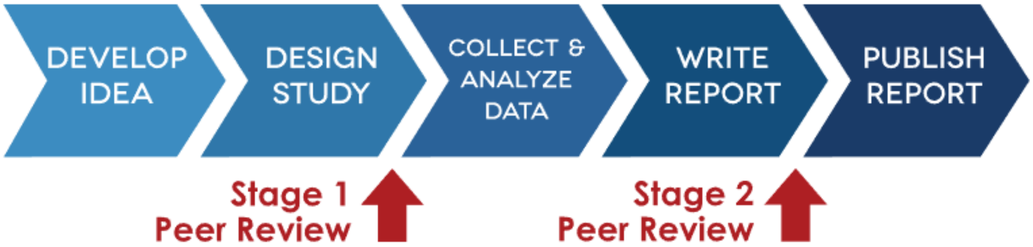


Figure 3.2: Registered Reports process. Reprinted from 'Center for Open Science', 2019,https://cos.io/rr/.

3.4 OSF

Credibility crisis led to the creation of the practice of open science which has been essential to share of protocols, data, and cods regarding the analysis of elements of studies.

Specifically, Open Science Framework (OSF) is a tool that promotes open, centralized work-flows by enabling capture of different aspects and products of the research life cycle, including developing a research idea, designing a study, storing and analyzing collected data, and writing and publishing reports or papers. It is developed and maintained by the Center for Open Science (COS), a nonprofit organization founded in 2013 that conducts research into scientific practice, builds and supports scientific research communities, and develops research tools and infrastructure to enable managing and archiving research

Particularly, OSF's mission is to increase openness, integrity, and reproducibility of research; particularly OSF has been defined as a game changer for those wanting to effectively share their research process in the spirit of collaboration. OSF helps research teams work on projects privately or make the entire project publicly accessible for broad dissemination. As a work-flow system, OSF enables connections to the many products researchers already use, streamlining their process and increasing efficiency. OSF is based on four pilasters:

- Structured projects: Manage files, data, code, and protocols in one centralized location and easily build custom organization for your project — no more trawling emails to find files or scrambling to recover lost data;
- Controlled access: Manage which parts of a project are public or private, making it easy to collaborate and share with the community or just your team;
- Enhanced work-flow: Connect your favorite third-party services like Dropbox or Google Drive, automate version control, preregister your research, share pre-prints.

- Dependable repository: OSF's preservation fund is sufficient for 50+ years of read access hosting at present costs.

Moreover, with OSF's work-flow and storage information, one can manage the entire project from one place.

OSF is one of the public repositories most largely used in psychology, although it's not the only one; for instance, another relevant repository is PsyArXiv ([PsyArXiv] (https://psyarxiv.com/)).

(OSF) indicates Registration Forms and Templates which are reported below. The site distinguish between *Available* and *Not Available* forms and templates on OSF.

Available on OSF:

1. OSF Prereg: Standard, comprehensive, and general purpose preregistration form. *Template*: Google Doc, OSF work-flow,R Markdown by Frederik Aust, R Markdown by James Bartlett;

2. Open-Ended Registration: Summary of registered work with a time-stamped snapshot of a research project. Use this one if you are registering a completed project with data or materials. *Template*: Word, GoogleDoc;

3. Qualitative Preregistration: Template for registering primarily qualitative work. *Template*: Word, GoogleDoc, FAQ;

4. AsPredicted Preregistration* form here Eight questions derived from content recommended by AsPredicted.org. *Template*: Word, GoogleDoc;

5. OSF-Standard Pre-Data Collection Registration: State whether data have been collected or viewed and other pertinent comments. Use this one if your pre-analysis plan is uploaded on OSF as a doc. *Template*: Word, GoogleDoc;

6. Replication Recipe (Brandt et al., 2013) *Pre-Registration*: Register a replication study with a series of questions regarding the original work. *Template*: Word, GoogleDoc;

7. Replication Recipe (Brandt et al., 2013) *Post-Completion*: Register a replication study after it has been conducted with questions regarding the outcomes of the replication. *Template*: Word, GoogleDoc;
8. Pre-Registration in Social Psychology (van 't Veer & Giner-Sorolla, 2016): Preregister a research study outlining the hypotheses, methods, and analysis plan. *Template*: Word, GoogleDoc, OSF;
9. Registered Report Protocol: Preregistration Register your protocol AFTER having been given "in-principle acceptance" from a Registered Report journal Word, GoogleDoc, OSF work-flow Secondary Data Preregistration* For preregistering a research project that uses an existing dataset. *Template*: OSF Page, Example, FAQ.

Not Available on OSF:

These forms are not available as guided work-flows on the OSF, but rather as template docs that you can fill out and register using the "Open-Ended Registration" form. We are always evaluating our guided work-flows and will likely include one or more of these in the future.

1. OLD "Qualitative Research Preregistration" This is an earlier version of the form that is now included on OSF. *Template*: OSF page;
2. Cognitive Modeling (Model Application)* Use when you wish to apply a cognitive model as a measurement tool to test hypotheses about parameters of the cognitive model. *Template*: OSF Page and Preprint;
3. fMRI Preregistration Template* This project provides a detailed preregistration template for fMRI studies and provides some guidance for common difficulties that can occur for fMRI preregistration projects. *Template*: OSF page;
4. Open Stats Lab and Project Tier* This preregistration template is geared towards researchers who have little experience with preregistering studies. *Template*: OSF

Page.

3.5 First studies on the value of Registered Reports

Recently, we have spotted increased attention towards Registered Reports.

As pointed out by Nosek & Lakens (2014), published journal articles are the primary means of communicating scientific ideas, methods, and empirical data. Sadly, not all ideas nor data get published. Particularly, Scheel (2021) stressed out that the practice of selectively publishing positive results, such as the ones which support the tested hypotheses, falsify the evidence for scientific claims: nowadays scientific divulgation culture identifies positive results more publishable than negative results. Consequently, we can assist a boost of the tendency to ignore replications and negative results, even at the expense of accuracy. As ultimate result, replications and negative results are infrequent in published literature.

Asendorpf et al. (2013) reported that in a comparison of publications in 18 empirical research areas, found rates of confirmed hypotheses ranging from 70% (space science) to 92% (psychology and psychiatry)(Asendorpf et al., 2013). The confirmation rate of 92% is above rates expected, given typical effect sizes and statistical power of psychological studies. Indeed, the rate seems to be inflated by selective non-reporting of non-confirmations as well as post hoc invention of hypotheses and study designs that do not subject hypotheses to the possibility of refutation (Asendorpf et al., 2013).

Additionally, a limpid practical example is provided by Scheel (2021) who sorted a random sample of hypotheses-tested studies from the standard psychological literature or Standard Reports (SRs) and compared the results with published Registered Reports (RRs). Scheel (2021) have analyzed the first hypotheses of each article; the results have demonstrated that almost the totality (96%) of positive results in SRs, while just nearly the half (44%) of positive results in RRs. The authors concluded that SRs might lead psychological scientists to miss out on many negative results from high-quality studies, which are available in the RR literature. It is essential to note that the absence of negative results is a serious threat.

Bakker et al. (2020b) freshly, published an article named *Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size*. In the paper, the authors investigated whether the statistical power of a study is higher when researchers are asked to make a formal power analysis before collecting data. Particularly, Bakker et al. (2020b) compared the sample size description from two sources:

1. a sample of pre-registrations created following the guidelines of the Center of Open Science and Preregistration Challenge (PCRs) and a sample of institutional review board (IRB) proposals from Tilburg School of Social Science, since both include recommendations to perform a power analysis;
2. a sample of pre-registrations created according to the guidelines for Open Science Framework Standard Pre-Data Collection Registrations (SPRs) in which no guidance on sample size planning is given.

The results suggested that PCRs and IRBs included more often (72%) sample size decisions based on power analyses than the SPRs (45%). Nonetheless, the sample size of PCRs and IRB proposals were not higher than the SPRs.

Additionally, only 20% of the power analyses hold a sufficient number of elements to wholly reproduce the results. Furthermore, uniquely 62% of these power analyses were related to the main hypotheses test in the pre-registration. Consequently, Bakker et al. (2020b) strongly asserted there is ample space for improvements in the quality of the registrations.

In conclusion, Scheel (2021) demonstrated that RRs practice enhances researchers to publish their articles on the basis of the quality of studies and not solely on their positive outcomes.

Fanelli (2010), in 2010, published the innovative article *Positive Results Increase Down the Hierarchy of the Science.* Initially Fanelli (2010) expressed the 200 years old hypothesis of a hierarchy of the sciences, where the physical sciences are at the top, the social sciences at the bottom and biological science in the middle. The order reflect the academic structure; however, according to the author, considering the *hardness* -defined by Fanelli (2010) himself as *the extent to which research questions and results are determined by data and theories as opposed to non-cognitive factors*- is tendentious. To determinate the hardness of sciences, the study analyzed nearly 2500 papers published which have declared they have tested their hypotheses in all the three disciplines: physical, biological and social sciences. Afterwards, the research equip ascertained the number of paper which reported a *positive* or *negative* support to their initial hypothesis. The result confirmed the Hierarchy Hypothesis: researches in *softer* sciences reported more positive outcomes when compared to the one at higher levels of the hierarchy, probably due to the fewer constrains and either conscious or unconscious biases which ultimately leaded to a methodology less fixed.

The result of Fanelli (2010) trial, suggested that both the nature of hypotheses and the methodological rigidity varies due to the type of science: depending on the complexity of the subject and the co-occurrence of other factors.

The paper underlines that positive results are frequent in social science fields, due to intrinsically laxity of methodology of these sciences and human biases are challenging to erase. Therefore, since social sciences are prone to incur in these biases, more attention ought to be dedicated.

More than ten years later, Scheel (2021) published the article *An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports* clearly inspired by the previous work of Fanelli (2010). During the last decade, as Fanelli (2010) brought to attention, social sciences are the most expose to psychological positive results, including psychology; therefore, psychology's scientists have been increasingly concerned about the degree of such distortion in their literature.

With the aim to prevent selective reporting, a new publication format has been proposed: Registered Reports (RRs), characterized by peer review; moreover publication decision takes place before results are known -thus are not influenced but positive or negative results.

Moving from these premises, Scheel (2021) compared the results in published RRs (N = 71 as of November 2018) with a random sample of hypothesis-testing studies from the standard literature (N = 152) in psychology. By analyzing the first hypotheses of each article, Scheel (2021) found 96% positive results in standard reports but only 44% positive results in RRs.

Scheel (2021) considered that a possible explanation of the enormous difference between positive results in standard reports and RRs are either the reduction of publication bias or Type I error inflation in the RR literature, or both.

Bakker et al. (2020c) conducted a similar study to Scheel (2021). Instead of focusing on the result of the study, in the preregistered study, Bakker et al. (2020c) in the evocative article called *Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size*, the authors investigated whether the statistical power of a study is higher when researchers are asked to make a formal power analysis before collecting data.

Bakker et al. (2020c) compared the sample size descriptions from two sources:

- A sample of pre-registrations created according to the guidelines for the Center for Open Science Preregistration Challenge (PCRs) and a sample of institutional review board (IRB) proposals from Tilburg School of Behavior and Social Sciences, which both include a recommendation to do a formal power analysis;
- A sample of pre-registrations created according to the guidelines for Open Science Framework Standard Pre-Data Collection Registrations (SPRs) in which no

guidance on sample size planning is given.

The authors discovered that the first source, such as PCRs and IRBs (72%) more often included sample size decisions based on power analyses when compared to the second sources being SPRs (45%). Despite the expectations, a more frequent inclusion of sample size failed to result in larger planned sample sizes: sample size of the PCRs and IRB proposals was not higher than the determined sample size of the SPRs. Additionally, Bakker et al. (2020c) pointed out usually power analyses in the registrations are conducted using G-power, assuming a medium effect size, $\alpha = .05$ and a power of 0.80. However, only the 20% of the power analyses had sufficient information to fully reproduce the results, plus only 62% of these power analyses relate to the main hypothesis test declared in the pre-registration. Finally, Bakker et al. (2020c) are convinced there is the change for utter improvements in the quality of the registration, and in the article they offered also numerous recommendations to practically implement those.

Fanelli (2010), Bakker et al. (2020c), Scheel (2021) assessed whether power analysis has been led on resulting sample size. Our work aims to deeply analyze sample size planning by including an evaluation of type and quantity of power-analysis conducted.

Chapter 4

Exploring OSF Registered Reports

In this chapter we present the exploratory study conducted. We took into account Registered Reports pre-registered on Open Science Framework (OSF), and we present an analysis of dataset at a descriptive level. We include critical evaluations of the results obtained.

4.1 Procedures and Data

We decided to utilize Open Science Framework (OSF) repository to gather data since OSF is one of the most popular public repositories used for psychology studies. OSF profoundly embraces the practice of open science, which has been essential to share of protocols, data, and codes regarding the analysis of elements of studies: OSF's mission is to increase openness, integrity, and reproducibility of research.

At the link (OSF) we found the Registered Reports (RRs) pre-registered. Among the 194 RRs available as of January 2022 (last view 22/01/2022) on OSF website, we have selected uniquely psychology inherent papers.

To successfully select psychological papers, we referred to Web of Science (WoS) classification.

The Web of Science (WoS; previously known as Web of Knowledge) is a paid-access platform that provides access to multiple databases that provide reference and citation data from academic journals, conference proceedings, and other documents in various academic disciplines. It was originally produced by the Institute for Scientific Information. Currently, it is owned by Clarivate (previously the Intellectual Property and Science business of Thomson Reuters).

Every journal, book and record covered by Web of Science core collection is assigned to at least one subject categories. Web of Science Categories field ([Web of Science] (https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html)) related to psychology are: Psychology, Applied; Psychology, Biological; Psychology, Clinical; Psychology, Developmental; Psychology, Educational; Psychology, Experimental; Psychology, Mathematical; Psychology, Multidisciplinary; Psychology, Psychoanalysis; Psychology, Social.

We read and coded the validity information for all measures used in the original and replication studies. Through the journal which publicated the Registered Report projects pre-registered on OSF, we individuate subject category (or categories). During this process, only articles published on journals which cover psychological categories have been selected to proceed to the next stage of analysis.

Among the 194 RRs available as of January 2022 (last view 22/01/2022) on OSF website, we recognize 46 Registered Reports (RRs) inherent to psychology.

4.2 Coding

We created a dataset containing the core information about the Registered Report Projects we analyzed.

We coded every Registered Report; in the dataset, we included Apa citations for every RRs of the dataset, as well as the authors and year of publication.

Additionally, we specified the Journal on where the Registered Report pre registered on OSF has been published, the Journal Impact Factor (IF), and Best Quartile (Q).

Furthermore, we relied on Wos Classification to assign psychological categories to each journal.

Moreover, we identified the type of software utilized for Power Analysis.

We assigned a typology of Sample Size Planning to every article. For every paper, we identified Sample Size Planning technique used and we assigned a number from 1 to 4 accordingly:

1. Absence of Sample Size declaration (NA);
2. Power analysis based on previous studies: authors decided to rely on psychological literature to determinate the sample size;
3. Power analysis based on a reasonable effect: authors conducted pilot studies based on psychological literature to determinate sample size;
4. Sequential analysis: In statistics, sequential analysis or sequential hypothesis testing is statistical analysis where the sample size is not fixed in advance. Instead data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed.

In addition, we assessed whether researchers relied on Cohen's Heuristic.

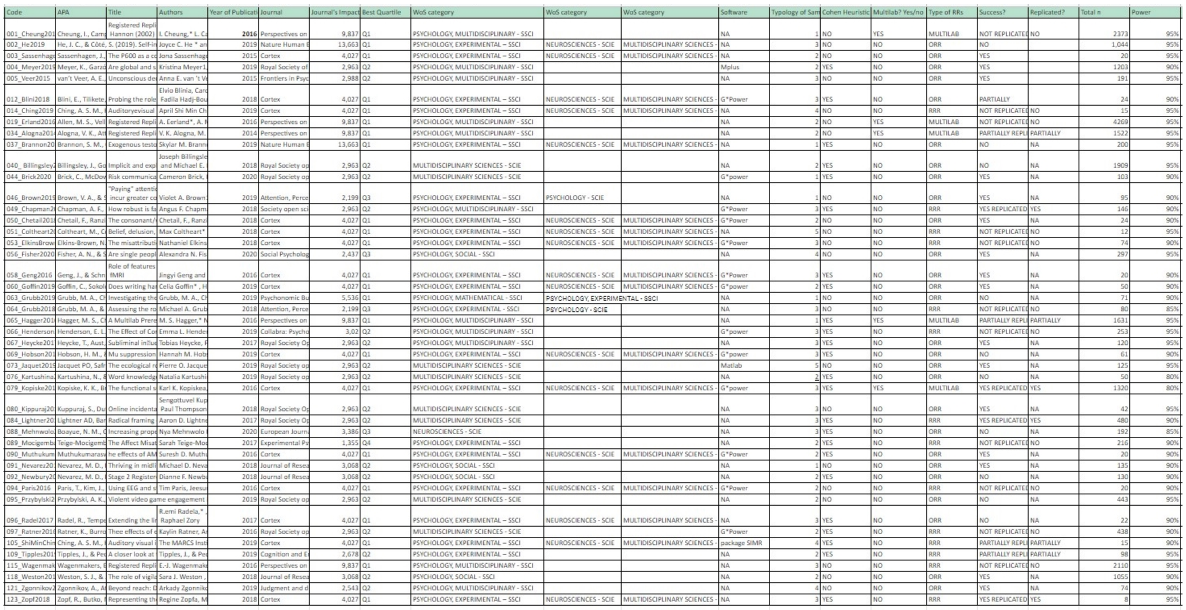
Later, we distinguished between: Original Registered Reports (ORRs), Replication Registered Reports (RRRs) and Multi-lab projects (Multi-lab).

Eventually, we referred to the First Hypothesis proposed by the researchers of each article to evaluate the Success, Unsuccess or Partial Success of every Registered Reports.

Finally, we reported the Sample Size and Power for each RR.

10 articles have been codified with a senior researcher to clarify possible doubts about coding.

For a visual representation of the Dataset refer to Figure 4.1



Code	APA	Title	Authors	Year of Publicati	Journal	Journal's Impact	Best Quartile	WoS category	WoS category	WoS category	Software	Typology of Sam	Cohen Heuristic	Multilab? Yes/no	Type of RRs	Success?	Replicated?	Total n	Power
001_Cheung201	Cheung, I., Camp	Registered Repli Hannon (2002)	I. Cheung,* L. Ca	**2016**	Perspectives on	9,837	Q1	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	1	NO	YES	MULTILAB	NOT REPLICATED	NO	2373	95%
002_He2019	He, J. C., & Côté, S. (2019). Self-in		Joyce C. He * an	2019	Nature Human E	13,663	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	3	NO	NO	ORR	NO		1,044	95%
003_Sassenhage	Sassenhagen, J.,	The P600 as a co	Jona Sassenhage	2015	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	2	NO	NO	ORR	YES		20	95%
004_Meyer2019	Meyer, K., Garzó	Are global and s	Kristina Meyer1,	2019	Royal Society of	2,963	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			Mplus	2	YES	NO	ORR	YES		1203	90%
005_Veer2015	van't Veer, A. E.,	Unconscious dec	Anna E. van 't Ve	2015	Frontiers in Psyc	2,988	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	3	NO	NO	ORR	YES		191	95%
012_Blini2018	Blini, E., Tilikete,	Probing the role	Elvio Blinia, Caro Fadila Hadj-Bou	2018	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	3	YES	NO	ORR	PARTIALLY		24	90%
014_Ching2019	Ching, A. S. M.,	Auditoryevisual	April Shi Min Ch	2019	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	4	NO	NO	RRR	NOT REPLICATED	NO	15	95%
019_Erland2016	Allen, M. S., Vell	Registered Repli	A. Eerland*, A. N	2016	Perspectives on	9,837	Q1	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	2	NO	YES	MULTILAB	NOT REPLICATED	NO	4269	95%
034_Alogna2014	Alogna, V. K., Att	Registered Repli	V. K. Alogna, M.	2014	Perspectives on	9,837	Q1	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	2	NO	YES	MULTILAB	PARTIALLY REPLI	PARTIALLY	1522	95%
037_Brannon20	Brannon, S. M.,	Exogenous testo	Skylar M. Branne	2019	Nature Human E	13,663	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	1	YES	NO	ORR	NO	NA	200	95%
040_ Billingsley2	Billingsley, J., Go	Implicit and expl	Joseph Billingsle and Michael E.	2018	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			NA	2	YES	NO	ORR	NO	NA	1909	95%
044_Brick2020	Brick, C., McDow	Risk communica	Cameron Brick,	2020	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			G*power	1	YES	NO	ORR	YES	NA	103	90%
046_Brown2019	Brown, V. A., & S	"Paying" attentio incur greater co	Violet A. Brown1	2019	Attention, Perce	2,199	Q3	PSYCHOLOGY, EXPERIMENTAL – SSCI	PSYCHOLOGY - SCIE		NA	1	NO	NO	ORR	YES	NA	95	90%
049_Chapman20	Chapman, A. F.,	How robust is fa	Angus F. Chapma	2018	Society open sci	2,963	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			G*Power	3	YES	NO	RRR	YES REPLICATED	YES	146	90%
050_Chetail2018	Chetail, F., Ranzi	The consonant/v	Chetail, F., Ranzi	2018	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	2	NO	NO	ORR	YES	NA	24	90%
051_Coltheart20	Coltheart, M., Ca	Belief, delusion,	Max Coltheart*	2018	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	5	NO	NO	RRR	NOT REPLICATED	NO	12	95%
053_ElkinsBrow	Elkins-Brown, N.	The misattributi	Nathaniel Elkins	2018	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	3	NO	NO	RRR	NOT REPLICATED	NO	74	90%
056_Fisher2020	Fisher, A. N., & S	Are single peopl	Alexandra N. Fis	2020	Social Psycholog	2,437	Q3	PSYCHOLOGY, SOCIAL - SSCI			NA	4	NO	NO	ORR	YES	NA	297	95%
058_Geng2016	Geng, J., & Schn	Role of features fMRI	Jingyi Geng and	2016	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	3	YES	NO	ORR	YES	NA	20	90%
060_Goffin2019	Goffin, C., Sokol	Does writing har	Celia Goffin* , H	2019	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*power	2	YES	NO	ORR	YES	NA	50	90%
063_Grubb2019	Grubb, M. A., Ch	Investigating the	Grubb, M. A., Ch	2019	Psychonomic Bu	5,536	Q1	PSYCHOLOGY, MATHEMATICAL - SSCI	PSYCHOLOGY, EXPERIMENTAL - SSCI		NA	1	NO	NO	ORR	NO	NA	71	90%
064_Grubb2018	Grubb, M. A., &	Assessing the ro	Michael A. Grub	2018	Attention, Perce	2,199	Q3	PSYCHOLOGY, EXPERIMENTAL - SSCI	PSYCHOLOGY - SCIE		NA	3	NO	NO	RRR	NOT REPLICATED	NO	80	85%
065_Hagger201	Hagger, M. S., Ch	A Multilab Prere	M. S. Hagger,* N	2016	Perspectives on	9,837	Q1	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	1	YES	YES	MULTILAB	PARTIALLY REPLI	PARTIALLY	1631	95%
066_Henderson	Henderson, E. L.	The Effect of Co	Emma L. Hende	2019	Collabra: Psycho	3,02	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			G*power	3	YES	NO	RRR	NOT REPLICATED	NO	253	95%
067_Heycke201	Heycke, T., Aust,	Subliminal influe	Tobias Heycke, F	2017	Royal Society Op	2,963	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	3	YES	NO	ORR	YES	NA	120	95%
069_Hobson201	Hobson, H. M., &	Mu suppression	Hannah M. Hobs	2019	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*power	3	YES	NO	ORR	NO	NA	61	90%
073_Jaquet2019	Jacquet PO, Safr	The ecological ro	Pierre O. Jacque	2019	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			Matlab	5	NO	NO	ORR	YES	NA	125	95%
076_Kartushina2	Kartushina, N., &	Word knowledge	Natalia Kartushi	2019	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			NA	2	YES	NO	ORR	NO	NA	50	80%
079_Kopiske201	Kopiske, K. K., Br	The functional s	Karl K. Kopiskea,	2016	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*power	3	YES	YES	MULTILAB	YES REPLICATED	YES	1320	80%
080_Kippuraj20	Kuppuraj, S., Du	Online incidenta	Sengottuvel Kup Paul Thompson	2018	Royal Society Op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			NA	3	NO	NO	ORR	YES	NA	42	95%
084_Lightner20	Lightner AD, Bar	Radical framing	Aaron D. Lightne	2017	Royal Society Op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			NA	3	YES	NO	RRR	YES REPLICATED	YES	480	90%
088_Mehnwolo	Boayue, N. M., C	Increasing prope	Nya Mehnwolo	2020	European Journa	3,386	Q3	NEUROSCIENCES - SCIE			NA	3	YES	NO	ORR	NO	NA	192	85%
089_Mocigemba	Teige-Mocigemb	The Affect Misat	Sarah Teige-Moc	2017	Experimental Ps	1,355	Q4	PSYCHOLOGY, EXPERIMENTAL – SSCI			NA	2	YES	NO	RRR	NOT REPLICATED	NO	216	90%
090_Muthukum	Muthukumarasw	he effects of AM	Suresh D. Muthu	2016	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	2	YES	NO	ORR	YES	NA	20	90%
091_Nevarez201	Nevarez, M. D.,	Thriving in midli	Michael D. Neva	2018	Journal of Resea	3,068	Q2	PSYCHOLOGY, SOCIAL - SSCI			NA	1	NO	NO	ORR	YES	NA	135	90%
092_Newbury20	Nevarez, M. D.,	Stage 2 Register	Dianne F. Newbu	2018	Journal of Resea	3,068	Q2	PSYCHOLOGY, SOCIAL - SSCI			NA	2	YES	NO	ORR	NO	NA	130	90%
094_Paris2016	Paris, T., Kim, J.,	Using EEG and s	Tim Paris, Jeesu	2016	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	G*Power	2	NO	NO	RRR	NOT REPLICATED	NO	20	90%
095_Przybylski2	Przybylski, A. K.,	Violent video game engagement		2019	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			NA	2	NO	NO	ORR	NO	NA	443	95%
096_Radel2017	Radel, R., Tempe	Extending the lin	R.emi Radela,* , Raphael Zory	2017	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	3	YES	NO	ORR	NO	NA	22	90%
097_Ratner2016	Ratner, K., Burro	Thee effects of e	Kaylin Ratner, Ar	2016	Royal Society op	2,963	Q2	MULTIDISCIPLINARY SCIENCES - SCIE			G*Power	2	YES	NO	RRR	NOT REPLICATED	NO	438	90%
105_ShiMinChin	Ching, A. S. M.,	Auditory visual i	The MARCS Insti	2019	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	package SIMR	4	YES	NO	RRR	PARTIALLY REPLI	PARTIALLY	15	90%
109_Tipples2019	Tipples, J., & Pec	A closer look at	Tipples, J., & Pec	2019	Cognition and E	2,678	Q2	PSYCHOLOGY, EXPERIMENTAL – SSCI			NA	2	YES	NO	RRR	PARTIALLY REPLI	PARTIALLY	98	95%
115_Wagenmak	Wagenmakers, E	Registered Repli	E.-J. Wagenmake	2016	Perspectives on	9,837	Q1	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	3	NO	NO	RRR	NOT REPLICATED	NO	2110	95%
118_Weston201	Weston, S. J., &	The role of vigila	Sara J. Weston ,	2018	Journal of Resea	3,068	Q2	PSYCHOLOGY, SOCIAL - SSCI			NA	2	NO	NO	ORR	YES	NA	1055	90%
121_Zgonnikov2	Zgonnikov, A., Ai	Beyond reach: D	Arkady Zgonnikc	2019	Judgment and d	2,543	Q2	PSYCHOLOGY, MULTIDISCIPLINARY - SSCI			NA	4	NO	NO	ORR	YES	NA	74	90%
123_Zopf2018	Zopf, R., Butko,	Representing th	Regine Zopfa, M	2018	Cortex	4,027	Q1	PSYCHOLOGY, EXPERIMENTAL – SSCI	NEUROSCIENCES - SCIE	MULTIDISCIPLINARY SCIENCES -	NA	3	YES	NO	RRR	YES REPLICATED	YES	8	95%

Figure 4.1: Dataset

4.3 Descriptive Analysis

4.3.1 Year of publication

Registered Reports is a procedure which has been introduced in 2012 by Chris Chambers, therefore it's still relatively new; on OSF, the first Registered Report has been publish on 2014. All the Psychological Registered Report projects inspected in this paper have been pre-registered on OSF between 2014 and 2020.

Precisely, a single article has been pre-registered in 2014 (Johnson et al. (2014a)). In 2015 two articles have been preregistered on OSF: Veer et al. (2015), and Sassenhagen & Bornkessel-Schlesewsky (2015).

The almost totality (93%) of psychology inherent Registered Reports have been pre-registered after 2015.

A sharp increase in the operation of pre-registration on OSF of Registered Report projects happened in 2016 when 9 articles have been pre-registered on OSF: the 19.57% of psychological RRs have been pre-registered on OSF in 2016.

In 2017, only five articles have been pre-registered on OSF; however a number of fourteen articles in 2018, and fifteen papers in 2019 have been pre-registered on OSF. A decrease happened in 2020 when only four articles have been pre-registered on OSF.

Table 4.1: The year of publication

Year	*Number*	*Percentage*
2014	1	2.17%
2015	2	4.34%
2016	9	19.5%
2017	5	10.86%
2018	14	30.43%
2019	15	32.60%
2020	4	8.69%

For a visual representation, refer to Figure 4.2.

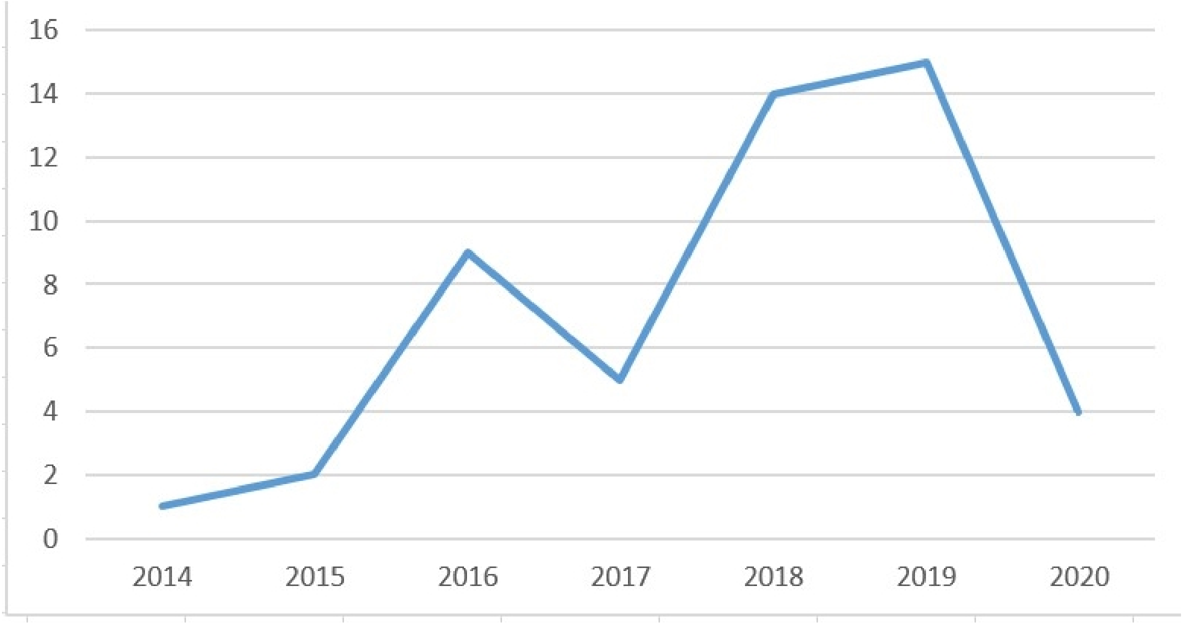


Figure 4.2: Number of Registered Reports pre-registered on OSF a year

The results might suggest that Registered Reports are a relatively new toos and it seems that their popularity didn't reach the peak yet. However, based on the findings, we are not able to make any prevision about the future developments of RRs.

4.3.2 Journals' Psychological Category, Impact factor and Best Quartile

The aim of this paper is to analyze psychological inherent Registered Reports pre-registered on OSF. To select the articles we referred to Web of Science classification.

Specifically, every journal covered by Web of Science core collection is assigned to at least one subject categories.

Web of Science Categories field ([Web of Science] (https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html)) related to psychology are: Psychology, Applied; Psychology, Biological; Psychology, Clinical; Psychology, Developmental; Psychology, Educational; Psychology, Experimental; Psychology, Mathematical; Psychology, Multidisciplinary; Psychology, Psychoanalysis; Psychology, Social.

WoS specifies the Impact Factor (IF) of a journal. IF is commonly used to evaluate the relative importance of a journal within its field and to measure the frequency with which the "average article" in a journal has been cited in a particular time period. Journal which publishes more review articles will get highest IFs.

Additionally, WoS assesses the best Quartile (Q) of a journal. The quartile is the ranking of a journal or paper definite by any database based on the impact factor (IF), citation, and indexing of that particular journal. It can divide into four different quadrants starting with Q1, Q2, Q3, and Q4.

Q index simply means quartile which is the ranking of any journal that belongs to a specific or particular field of discipline and also known as the parameter of measuring or ranking of a journal.

- Quartile 1 (Q1): The first position of the top 25% of journals in a particular category are placed in this category (top 25%);

- Quartile 2 (Q2): The middle-high position subsequent occupied by 25% Journal after quartile 1 fall under this category (between top 25% to 50%);
- Quartile 3 (Q3): The middle-low position next 25% Journal title after Q2 fall under this category (between 50% to 75%);
- Quartile 4 (Q4): The last or lowest position following 25% Journal title of a selected field will fall under this category (between 75% to 100%).

In our dataset, we observed that: the majority (32.6%) of Registered Reports are published on *Cortex*, followed by *Royal Society Open Science* (13.04%). Interesting, we noted that *Cortex* has an IF=4,027 and Q1 while *Royal Society Open Science* IF=2,963 and Q2, therefore both are importnat journals (high IF) and high rank (Q1 and Q2).

However, the Registered Reports taken into account in the present exploratory study, *Nature Human Behaviour* is the journal with the highest Impact Factor (IF= 13,663) and it has Q1, while *Collabra: Psychology*, is the journal with the lowest IF (IF= 3,02) and it has Q2. A single RR has been published on *Nature Human Behaviour* and only one RR has been published on *Collabra: Psychology.*

In regards of the Best Quartile, *Experimental Psychology* is the only journal with Q4. It's worth noting that only one of the 46 Registered Report projects considered in the present exploratory study has been published on this journal. 23 (50%) Registered Report projects have been published on a journal with Q1; while 18 (39.13%) Registered Reports have been published on a journal with Q2 and solely 4 (8.69%) Registered Reports have been published on a journal with Q3. Therefore, 89.13% of Registered Report projects taken into account in the present experimental study have been published on journals with Q1 or Q2.

This result could suggest that on average, RRs are published on high ranked journals.

For what concerns the psychological categories listed on WoS, we assessed that the most frequently found in Registered Report projects is Psychology, experimental; followed by Psychology, multidisciplinary and Psychology, social.

Other seven categories, such as: Psychology, Applied; Psychology, Biological; Psychology, Clinical; Psychology, Developmental; Psychology, Educational; Psychology, Mathematical; Psychology, Psychology, Psychoanalysis aren't remarkably present in the RRs pre-registered on OSF that we analyzed.

To sum it up, it is feasible to state that among ten psychological categories listed on WoS, three are the more likely to be found in RRs pre-registered on OSF, whereas the other seven psychological categories aren't significantly portrayed. A possible explanation could be that Psychology, experimental is a branch of psychology where researchers usually are able to break up complex hypotheses in simplified statements which are easier to test. Therefore, Psychology, experimental usually proposes simplified questions of research and researchers can pre-registered the project with less difficulties when compared to other psychology categories such as Psychology, Psychoanalysis.

4.3.3 Typology of Registered Report

We have identified three types of Registered Reports:

- Multi-lab replication projects (Multi-lab): In a Multi-lab replication project, multiple teams of investigators team up to all run the same study concurrently at different research sites to replicate an original study. It's important to note that Multi-lab projects might also be original studies. However, in the present paper, all the Multi-lab projects are replications of original studies.
- Registered Report replication project (RRR): a single team of investigators run a replication of an original study.
- Original Registered Report project (ORR) : a single team of investigators run an original study.

Briefly, Multi-lab and RRR projects are a replication of original studies, whereas ORR projects are original studies.

Among the psychological Registered Reports (RRs) pre-registered on Open Science Framework (OSF), as on January 2022, we found:

- 5 (10.86%) Multi-lab (Multi-lab), all the Multi-lab studies are replications of original studies e.g. Hagger et al. (2016a);
- 14 (30.43%) Replication Registered Reports (RRR): such as papers which replicated an original study e.g. Grubb & Li (2018);
- 27 (58.69%) Original Registered Reports (ORR): such as articles which have conducted an original study e.g. Grubb et al. (2019a).

Table 4.2: Type of Registered Reports

RRs	*Number*
Multi-lab	5
RRR	14
ORR	27

For a visual representation, refer to Figure 4.3.

According to the results obtained, it is feasible to state that majority of RRs are original projects (ORR), a third are Replication Registered Reports and few are Multi-lab projects. A possible explanation of these results could be that Multi-lab projects are difficult to run because more laboratories are involved, and a relevant number of researchers need to coordinate their work whereas RRR and ORR are easier to run.

4.3.4 Type of Software Used for Power Analysis

A researcher can rely on software to calculate power analysis.

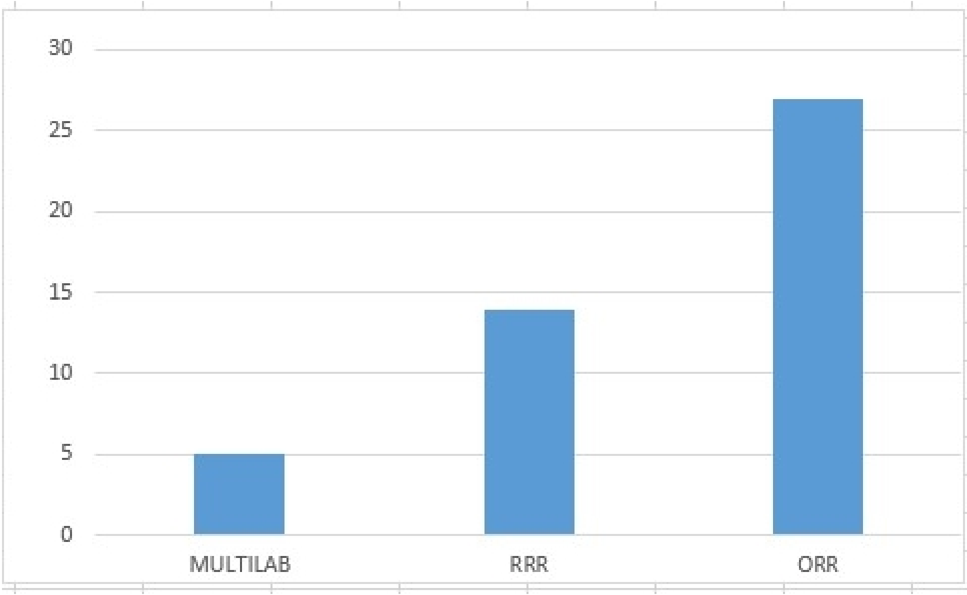


Figure 4.3: Types of Registed Reports pre-registered on OSF

In our survey, we observed that:

- 13 (28.26%) RRs relied on G-power, e.g. Muthukumaraswamy et al. (2016).

G-Power is a tool to compute statistical power analyses for many different tests. G-Power can also be used to compute effect sizes and to display graphically the results of power analyses.

- Ching et al. (2019) used Package 'simr.'

Package 'simr' calculates power for generalized linear mized models, using simulation. Described in Green and MacLeod, 2016.

- Jacquet et al. (2019) used MATLAB.

MATLAB it is the abbreviation of "MATrix LABoratory." MATLAB have been created at the end of 1970 by Cleve Moler. MATLAB combines a desktop environment tuned for iterative analysis and design processes with a programming language that expresses matrix and array mathematics directly. It includes the Live Editor for creating scripts that combine code, output, and formatted text in an executable notebook.

- Meyer et al. (2019) used Mplus.

Mplus offers researchers a wide choice of models, estimators, and algorithms in a program that has an easy to use interface and graphical displays of data and analysis results. Mplus is a latent variable modeling program with a wide variety of analysis capabilities, including: Exploratory factor analysis, complex survey data analysis, Bayesian analysis. Described in 1998

- 30 (65.22%) Register Report Projects didn't specify the software used to run Power Analyses e.g. Tipples & Pecchinenda (2019).

Table 4.3: Software used for Power Analysis

Software	*Number*	*Percentage*
G-power	13	28.26%
Package 'simr'	1	2.17%
MATLAB	1	2.17%
Mplus	1	2.17%
NA	30	65.22%

For a visual representation, refer to Figure 4.4.

We could state that the majority of RRs didn't specify the software used to run Power Analysis. Despite this, the most popular software used by researchers for Power Analysis is G-Power.

4.3.5 Success

What does it mean that a Registered Report is successful?

To answer the present question, we have to precise that there is a difference between Original Registered Reports projects when compared to Replication Registered Reports

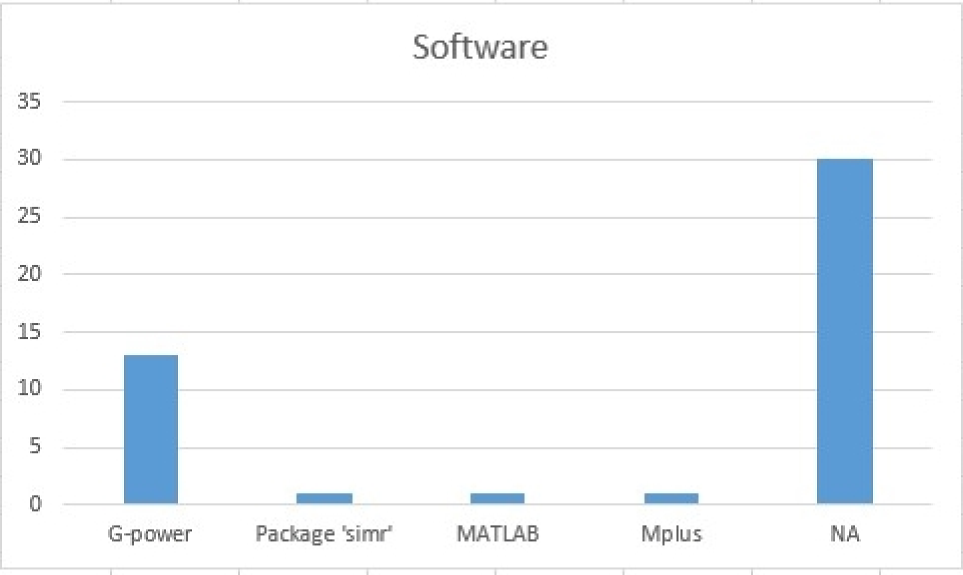


Figure 4.4: Software used to run Power Anlyses in Registerd Reports pre-registered on OSF

and Multi-lab projects.

In the case of ORR studies, we have three options:

- the ORR is successful meaning the result confirmed the first hypothesis proposed;
- the ORR is not-successful when it the result didn't confirm the first hypothesis proposed;
- the ORR is partially successful when the effect found is smaller than what assumed in the first hypothesis.

For Replication Registered Reports and Multi-lab projects, we have three additional options:

- RRs and Multi-labs replicated successfully the first hypothesis proposed;
- RRs and Multi-labs didn't replicate successfully the first hypothesis proposed;
- RRs and Multi-labs are partially successful when the effect found is smaller than what assumed in the first hypothesis.

The difference between ORR, RRS and Multi-lab is that in the first case, the first hypothesis is an original statement, whereas in the latter cases, RRs and Multi-lab

replicated original studies to confirm or not the original hypotheses.
It's essential to note that the absence of negative results is a serious threat, because it usually implies a high rate of false positive results.

We observed that among ORR projects:

- 16 ORR successfully confirmed their hypotheses e.g. Sassenhagen & Bornkessel-Schlesewsky (2015);
- 10 ORR didn't confirm their hypotheses e.g. He & Côté (2019);
- 1 ORR partially replicated e.g. Blini et al. (2018).

Table 4.4: ORRs Success

ORRs	*Number*	*Percentage*
Success	16	59.25%
Unsuccess	10	37.03%
Partial success	1	3.7%

Among replications, which included both Multi-lab projects and RRRs:

- 4 replicated;
- 11 didn't replicated;
- 4 partially replicated.

Table 4.5: RRs and Multi-labs Success

ReplicationSuccess	*Number*	*Percentage*
Success	4	21.05%
Unsuccess	11	57.89%%

ReplicationSuccess	*Number*	*Percentage*
Partial success	4	21.05%

On 19 replication projects, only 21.05% replicated successfully.

Specifically, among Multi-lab projects:

- 1 Multi-lab successfully replicate Kopiske et al. (2016a);
- 2 Multi-lab didn't replicate Cheung et al. (2016);
- 2 Multi-lab partially replicated Hagger et al. (2016b).

Out of 5 Multi-lab projects, 20% replicated successfully.

Table 4.6: Multi-labs Success

Multi-labs	*Number*	*Percentage*
Success	1	20%
Unsuccess	2	40%
Partial success	2	40%

These results might suggest that the rate of success in Original Registered Report Project is 59.26%; whereas the rate of success of Replication Projects (Multi-lab and RRR), is 21.05%.

However, by considering Multi-lab projects separately, the rate of success is 20%.

The result advocates that Replication Projects (i.e. Multi-lab and RRR) are less likely to find a significant effect when compared to Original Replication Projects. A possible explanation is that Replication Projects

We want to stress again that, although it seems counterintuitive, the absence of negative results in publishing literature is a serious threat, since it implies a higher likelihood to find false positive results.

Specifically, a study conducted by Asendorpf et al. (2013) found rates of confirmed hypotheses of 92% in psychology and psychiatry. The confirmation rate of 92% is above rates expected, given typical effect sizes and statistical power of psychological studies. Indeed, the rate seems to be inflated by selective non-reporting of non-confirmations as well as post hoc invention of hypotheses and study designs that do not subject hypotheses to the possibility of refutation (Asendorpf et al., 2013). RRs projects are a useful tool to avid to miss out on many negative results from high-quality studies, which are available in the RR literature.

Another research conducted by Scheel (2021) sorted a random sample of hypothesis-tested studies from the standard psychological literature of Standard Reports (SRs) and compared the results with published Registered Reports (RRs). Scheel (2021) have analyzed the first hypothesis of each article; the results have demonstrated almost the totality (96%) of positive results in SRs, while just nearly the half (44%) of positive results in RRs. The authors concluded that SRs might lead psychological scientists to miss out on many negative results from high-quality studies, which are available in the RR literature.

Scheel (2021) demonstrated that RRs practice enhances researchers to publish their articles on the basis of the quality of the study and not solely on its positive outcome as happened for SRs: by analyzing the first hypothesis of each article, Scheel (2021) found 96% positive results in standard reports but only 44% positive results in RRs.

Scheel (2021) considered that a possible explanation of the enormous difference between positive results in standard reports and RRs are either the reduction of publication bias or Type I error inflation in the RR literature, or both.

Fanelli (2010), Bakker et al. (2020c), Scheel (2021) assessed whether power analysis has been led on resulting sample size.

The present work supports the result obtained by the authors aforementioned: Registered Reports could contribute to reduce the presence of false positive results in the literature.

4.3.6 Sample Size

The Sample Size used in the Registered Reports we analyzed had a Median=111.5, SD=843.57, MIN=8, MAX=4269.

Table 4.7: RRs Statistical indexes

IndexRR	*Value*
Median	111.5
SD	843.57
MIN	8
MAX	4269

Taking into account only Multi-lab projects, Median is Mean=1631, with a SD=655.84, MIN=1320, MAX=4269.

Table 4.8: Multi-lab Statistical indexes

IndexMULTILAB	*Value*
Median	1631
SD	655.84
MIN	1320
MAX	4269

Based on these results, we could state that the variability of the number of participants recruited profoundly varies. However, considering Multi-lab projects alone, it is possible to observe that there is less variability among them when compared to the general RRs. A possible explanation could be that Multi-lab projects are more homogeneous since these projects imply a coordination of multiple labs.

4.3.7 Power

According to Neyman-Pearson, the power of a hypothesis test is the probability that test correctly rejects the null hypothesis when the alternate hypothesis is true. For example, a power of 0.9, means that 90% of the time one would get a statistically significant result while in 10% of the cases, results would not be statistically significant.

- 20 (43.48%) are the articles which used 95% of power e.g. Johnson et al. (2014a);
- 22 (47.82%) are the articles which used 90% of power e.g. Meyer et al. (2019);
- 2 (4.35%) are the articles which used 85% of power e.g. Grubb et al. (2019b);
- 2 (4.35%) are the articles which used 80% of power e.g. Kartushina & Mayor (2019).

Table 4.9: Power used by Registered Reports pre-registered on OSF

Power	*Number*	*Percentage*
95%	20	43.48%
90%	22	47.82%
85%	2	4.35%
80%	2	4.35%

For a visual representation, refer to Figure 4.5.

Based on the result, we can observe that the majority of Registered Reports analyzed in the present paper used a power of 90% (22 articles) and 95% (20 articles), whereas we individuated only 4 articles in total which used a power of 85% or 80%.

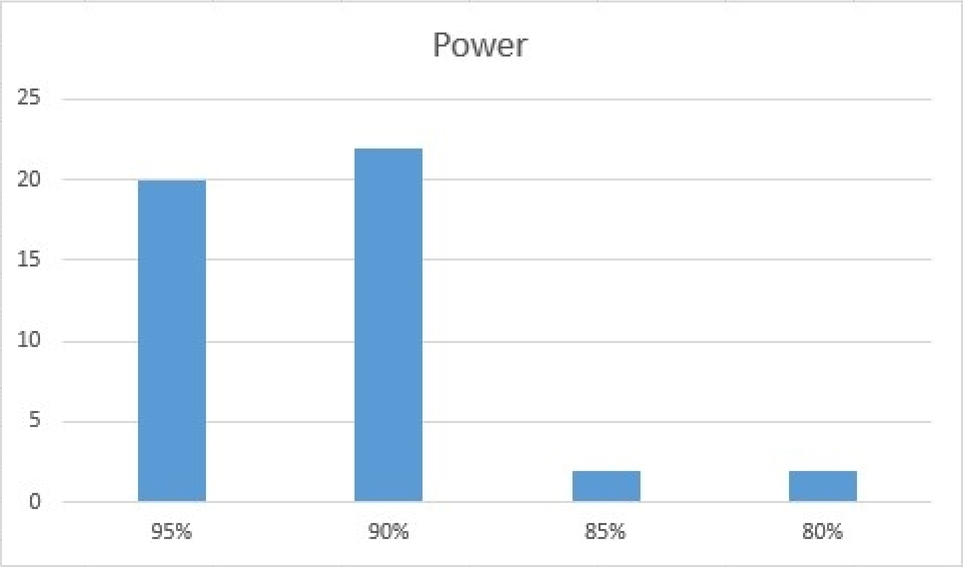


Figure 4.5: Power used by Registered Reports pre-reigstered on OSF

Based on these results, we could express that most of the RRs used a high power of 95% or 90%.

4.3.8 Power Analysis

For every paper, we identified Sample Size Planning technique used and we assigned a number from 1 to 4 accordingly:

1. Absence of Sample Size declaration (NA);
2. Power analysis based on previous studies: authors decided to rely on psychological literature to determinate the sample size;
3. Power analysis based on a reasonable effect: authors conducted pilot studies based on psychological literature to determinate sample size;
4. Sequential analysis: In statistics, sequential analysis or sequential hypothesis testing is statistical analysis where the sample size is not fixed in advance. Instead data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed.

We carefully inspected the 46 Registered Reports pre-registered on OSF (as on January 2022). We ascertained that:

- 7 (15.21%) RRs didn't declare how they conducted power analysis e.g. Nevarez et al. (2018);
- 16 (34.78%) RRs based their power analysis on statistical literature e.g. Przybylski & Weinstein (2019);
- 17 (36.95%) RRS conducted a pilot study and therefore based their power analysis on a reasonable effect e.g. Zopf et al. (2018);
- 6 (13.04%) RRs relied on a sequential analysis e.g. Zgonnikov et al. (2019).

Table 4.10: RRs Sample Size Technique

RRSSTechnique	*Number*	*Percentage*
1	7	15.21%
2	16	34.78%
3	17	36.95%
4	6	13.04%

Later, we decided to focus solely on Multi-labs.

Among Multi-lab projects:

- In two of them there is an absence of details about how power analysis has been conducted (Hagger et al. (2016b) and Johnson et al. (2014b) (both published on Perspectives on Psychological Science));
- Two Multi-lab projects relied on statistical literature to conduct its power analysis (Alogna et al. (2014) and Eerland et al. (2016) (both published on Perspectives on Psychological Science));

- One Multi-lab project based its power analysis on reliable effects (Kopiske et al. (2016b) (published on Cortex)).

Table 4.11: Multi-labs Sample Size Technique

MULTILABSSTechnique	*Number*	*Percentage*
1	2	40%
2	2	40%
3	1	20%
4	0	0%

This result suggests that Multi-lab projects don't necessarily guarantee a thoughtful Power Analysis.

Furthermore, we analyzed the use of Cohen's Heuristic to determine effect size. We puzzled out that among the 46 RRs in our dataset:

- 24 (52.17%) articles relied on Cohen's Heuristic to determine effect size, whereas 22 registered Reports didn't use Cohen's Heuristic;

10 of the RRs have been inspected and verified by an expert.

Table 4.12: Use of Cohen Heuristic (CH) in Registered Reports

CH	*Number*	*Percentage*
Application of Cohen's Heuristic	24	52.17%
Not application of Cohen's Heuristic	22	47.83%

We should note that the blind use of Cohen's Heuristic might be a serious threat in regards of scientific advancement, because of the absence of any justification for Effect

Size.

Focusing only on Multi-lab projects:

- two applied Cohen's Heuristic for Effect Size;
- three didn't use Cohen's Heuristic to determine the Effect Size.

Table 4.13: Use of Cohen Heuristic in Multi-lab projects (CHM)

CHM	*Number*	*Percentage*
Application of Cohen's Heuristic	2	40%
Not application of Cohen's Heuristic	3	60%

Even in the specific case of Multi-lab studies, the rate of studies which applied Cohen's Heuristic is similar to the one of the general RRs. Relying on Cohen's Heuristic might simplification since the researchers don't justify how they determine the Effect Size. What we could observe is that RRs don't guarantee a well planned Power Analysis. Specifically, the 40% of RRs didn't specify the Sample Size technique used, which might be a serious threat since the procedure is obscure. However, it's also worth noting that 36.95% of RRs run a well planned power analysis basing on psychology literature as well as pilot studies.

In this chapter we evaluated at a descriptive level the most common practices to Plan for Sample Size, while we individuated points of strength and weakness.

In this chapter, we analyzed the year of publication noting that RRs haven't probably reached their peak of popularity yet, since they have been proposed in 2012, just ten years ago. Furthermore, we assessed that many RRs pre-registered on OSF have been published on high ranked and trustworthy journals with high IF and Q. However, the psychological categories present are not inclusive, only three are portraited and

Psychology, Experimental is the most likely to be pre-registered. We distinguished among three types of Registered Reports: ORR, RRR and Multi-lab; we noted that Multi-lab are few, probably because many labs have to collaborate and be coordinate anf this requires a lot of effort. We observed that the variability of the number of participants recruited profoundly varies. However, considering Multi-lab projects alone, it is possible to observe that there is less variability among them when compared to the general RRs. A possible explanation could be that Multi-lab projects are more homogeneous since these projects imply a coordination of multiple labs. Additionally, the majority of Registered Reports analyzed in the present paper used a power of 90% (22 articles) and 95% (20 articles), whereas we individuated only 4 articles in total used a power of 85% or 80%.: most of the RRs used a high power of 95% or 90%. Finally, the 40% of RRs didn't specify the Sample Size technique used, which might be a serious threat since the procedure is obscure. However, it's also worth noting that 36.95% of RRs run a well planned power analysis basing on psychology literature as well as pilot studies.

Overall, RRs may be considered a procedure which can enhance the quality of papers, although there is room for improvements.

Future improvements should focus on promoting a more thoughtful Sample Size planning while enhancing the application of pilot studies based on psychological literature to determine an adequate sample size.

Chapter 5

General comments

Psychology faced an out of ordinary credibility crisis. As Ioannidis (2005) pointed out, a considerable number of research's findings are unclear, unreliable or cannot be replicated. Therefore, the low replication rate of research findings in psychology and the consequently credibility crisis in the published results increased the awareness of the problematic issues in the literature.

However, the crisis, as a side effect, promoted the development of new practices aimed to enhance rigorousness, reproducibility, and transparency.

Among these practices, pre-registration and pre-registered report, the creation of online long-term data repositories where to share materials and data (i.e., Open Science Frameworks) and the increased awareness that during the research process multiple factors must be taken into consideration (e.g., power, sample size, plausible effect size).

In the present work the aim is to provide a small contribution analyzing whether Replication Reports can enhance Sample Size Planning and consequently Power Analysis. Specifically, we conducted an exploratory study by taking into account the psychological RRs pre-registered on OSF as of January 2022 to test the question of research.

As described in the second chapter, on one hand Statistical Power Analysis utilizes the mathematical connection among the four quantities (i.e. *power*, α, n, and *ES*) in statistical inference, since the relationships between them allows the expert to determine the fourth element when the other three are set (Gigerenzer et al., 2004). In this scenario, it is possible to minimize errors simultaneously to increase statistical power. Because replicable result are more likely when power is high, the core process is identifying the factors that increase statistical power; that is, for any α level, statistical power increases as effect size and sample size increase. Afterwards, we described Design Analysis as a tool which allows to evaluate two inferential risks, namely Type M and Type S errors and can be useful both during the research process and during results' evaluation.

These threats, are usually combined with the presence of publication bias which are described in the third chapter. The latter favors significant result over null or negative findings, which could lead to misleading and unreliable results. Indeed, Questionable Research Practices (QRPs) are methodological and statistical practices that bias the scientific literature and affect credibility and reproducibility of research findings.

Registered Reports (RRs) have been proposed in 2012 as a tool to free researchers from the pressure to engage in the counterproductive practices aforemetioned, thereby breaking the cycle that perpetuates bias and absence of reproducibility. The RR model originates from the simple philosophy that to defeat the distorting effects of outcome bias on science, we must focus on the process and blind the evaluation of science to research outcomes. RRs are therefore a way of implementing a review procedure which is peer-reviewed before collecting the data and are published independently of the final results. In this way, registered reports help to prevent low statistical power, selective reporting of results, and publication bias (Bakker et al., 2020b).

Many RRs are pre-registered on Open Science Frameworks(OSF), a tool that promotes open, centralized work-flows by enabling capture of different aspects and products of the research life cycle, including developing a research idea, designing a study, storing and analyzing collected data, and writing and publishing reports or papers. Particularly, OSF's mission is to increase openness, integrity, and reproducibility of research.

In the fourth chapter, we presented our exploratory study, which aimed to assess whether Registered Reports enhance Sample Size Planning.

We conducted a systematic review of the measures used original and replication studies from the Reproducibility Project Psychology (Open Science Framework). We utilized OSF repository to gather data since OSF is one of the most popular public repositories used for psychology studies. Afterwards, we proposed a Descriptive Analysis of the elements of the dataset. We considered various elements such as: Year of publication, Journals' Psychological Category, Impact factor and Best Quartile. Additionally we have identified three types of Registered Reports (i.e. Original Registered Reports, Registered Report Replication projects and MULTILAB projects), we pointed out the type of Software Used for Power Analysis (e.g. G-power, Package 'simr,' MATLAB and MPLUS), the Power used, additionally we tracked the rate of false positive results.

We noted that Registered Reports reduce false positive results. This finding is in line with the findings of Fanelli (2010), Bakker et al. (2020c), Scheel (2021).

In conclusion, we believe that Registered Reports improve Sample Size Planning, therefore Registered Reports are useful to diminish potential risks associated with study results. However, we discovered that not all the Registered Report projects published on OSF run well planned Power Analyses, therefore we suggest to upgrade the preregistration criterion, in this way we expect a reduction of obscurities and unclarity.

Limits

The present exploratory study inspected 46 Registered Report projects pre-registered only Open Science Framework (OSF), without taking into account other online repositories, therefore it could not be considered inclusive. We focused on Power Analysis with a specific focal point on Sample Size Planning to enhance the Power of a study, however Effect Size, and Alpha level are worthy of further analysis.

Further developments

Further studies might embrace PsyArXiv as it a relevant psychology repository. During the analysis, we focus mainly on Sample Size Planning, however, further studies might focus on Effect Size or Alpha level since those are other elements which can enhance the power of a study.

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., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., … 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., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology*, *10*, 2893. https://doi.org/10.3389/fpsyg.2019.02893

Amaratunga, D., Baldry, D., Sarshar, M., & Newton, R. (2002). Quantitative and qualitative research in the built environment: Application of “mixed” research approach. *Work Study*, *51*(1), 17–31. https://doi.org/10.1108/00438020210415488

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of*

Personality, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Bakker, M., Veldkamp, C. L. S., Akker, O. R. van den, Assen, M. A. L. M. van, Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020b). Recommendations in preregistrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE*, *15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Bakker, M., Veldkamp, C. L. S., Akker, O. R. van den, Assen, M. A. L. M. van, Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020a). Recommendations in preregistrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE*, *15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Bakker, M., Veldkamp, C. L. S., Akker, O. R. van den, Assen, M. A. L. M. van, Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020c). Recommendations in preregistrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE*, *15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Bertoldo, G. (2019). *Dealing with the replication crisis in psychological science: The contribution of Type M and Type S errors* [Preprint]. Thesis Commons. https://doi.org/10.31237/osf.io/w63h7

Bertoldo, G., Zandonella Callegher, C., & Altoè, G. (2020). *Designing Studies and Evaluating Research Results: Type M and Type S Errors for Pearson Correlation Coefficient* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/q9f86

Blini, E., Tilikete, C., Farnè, A., & Hadj-Bouziane, F. (2018). Probing the role of the vestibular system in motivation and reward-based attention. *Cortex*, *103*, 82–99. https://doi.org/10.1016/j.cortex.2018.02.009

Brockett, R. W., & Mesarović, M. D. (1965). The reproducibility of multivariable systems. *Journal of Mathematical Analysis and Applications*, *11*, 548–563. https:

//doi.org/10.1016/0022-247X(65)90104-6

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

Calin-Jageman, R. J., & Caldwell, T. L. (2014). Replication of the Superstition and Performance Study by. *Social Psychology*, *45*(3), 239–245. https://doi.org/10.1027/1864-9335/a000190

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., … Yong, J. C. (2016). Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. https://doi.org/10.1177/1745691616664694

Ching, A. S. M., Kim, J., & Davis, C. (2019). Auditoryvisual integration during nonconscious perception. *Cortex*, *117*, 1–15. https://doi.org/10.1016/j.cortex.2019.02.014

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

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Elst, H. van. (2019). Foundations of Descriptive and Inferential Statistics. *arXiv:1302.2525 [Stat]*. https://doi.org/10.13140/RG.2.1.2112.3044

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

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

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 *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n21

Given, L. M. (Ed.). (2008). *The Sage encyclopedia of qualitative research methods.* Sage Publications.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341). https://doi.org/10.1126/scitranslmed.aaf5027

Grubb, M. A., Christensen, G., & Albanese, J. (2019b). Investigating the role of exogenous cueing on selection history formation. *Psychonomic Bulletin & Review*, *26*(4), 1282–1288. https://doi.org/10.3758/s13423-019-01591-z

Grubb, M. A., Christensen, G., & Albanese, J. (2019a). Investigating the role of exogenous cueing on selection history formation. *Psychonomic Bulletin & Review*, *26*(4), 1282–1288. https://doi.org/10.3758/s13423-019-01591-z

Grubb, M. A., & Li, Y. (2018). Assessing the role of accuracy-based feedback in value-driven attentional capture. *Attention, Perception, & Psychophysics*, *80*(4), 822–828. https://doi.org/10.3758/s13414-018-1494-y

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016b). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016a). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

He, J. C., & Côté, S. (2019). Self-insight into emotional and cognitive abilities is not related to higher adjustment. *Nature Human Behaviour*, *3*(8), 867–884. https://doi.org/10.1038/s41562-019-0644-0

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

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

Jacquet, P. O., Safra, L., Wyart, V., Baumard, N., & Chevallier, C. (2019). The ecological roots of human susceptibility to social influence: A pre-registered study investigating the impact of early-life adversity. *Royal Society Open Science*, *6*(1), 180454. https://doi.org/10.1098/rsos.180454

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014b). Does Cleanliness Influence Moral Judgments?: A Direct Replication of. *Social Psychology*, *45*(3), 209–215. https://doi.org/10.1027/1864-9335/a000186

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014a). Does Cleanliness Influence Moral Judgments?: A Direct Replication of. *Social Psychology*, *45*(3), 209–215. https://doi.org/10.1027/1864-9335/a000186

Kartushina, N., & Mayor, J. (2019). Word knowledge in six- to nine-month-old Norwegian infants? Not without additional frequency cues. *Royal Society Open Science*, *6*(9), 180711. https://doi.org/10.1098/rsos.180711

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

Kopiske, K. K., Bruno, N., Hesse, C., Schenk, T., & Franz, V. H. (2016a). The functional subdivision of the visual brain: Is there a real illusion effect on action? A multi-lab replication study. *Cortex*, *79*, 130–152. https://doi.org/10.1016/j.cortex.2016.03.020

Kopiske, K. K., Bruno, N., Hesse, C., Schenk, T., & Franz, V. H. (2016b). The functional subdivision of the visual brain: Is there a real illusion effect on action? A multi-lab replication study. *Cortex*, *79*, 130–152. https://doi.org/10.1016/j.cortex.2016.03.020

Lakens, D. (2021). *Sample Size Justification* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Concep-*

tual Analysis [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jbh4w

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

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). *Research Preregistration 10.* 8.

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

Meyer, K., Garzón, B., Lövdén, M., & Hildebrandt, A. (2019). Are global and specific interindividual differences in cortical thickness associated with facets of cognitive abilities, including face cognition? *Royal Society Open Science*, *6*(7), 180857. https://doi.org/10.1098/rsos.180857

Muthukumaraswamy, S. D., Routley, B., Droog, W., Singh, K. D., & Hamandi, K. (2016). The effects of AMPA blockade on the spectral profile of human early visual cortex recordings studied with non-invasive MEG. *Cortex*, *81*, 266–275. https://doi.org/10.1016/j.cortex.2016.03.004

Nevarez, M. D., Morrill, M. I., & Waldinger, R. J. (2018). Thriving in midlife: The roles of childhood nurturance and adult defense mechanisms. *Journal of Research in Personality*, *74*, 35–41. https://doi.org/10.1016/j.jrp.2018.01.002

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Przybylski, A. K., & Weinstein, N. (2019). Violent video game engagement is not associated with adolescents' aggressive behaviour: Evidence from a registered report. *Royal Society Open Science*, *6*(2), 171474. https://doi.org/10.1098/rsos.171474

Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex*, *66*, A3–A20. https://doi.org/10.1016/j.cortex.2014.12.019

Scheel, A. M. (2021). *An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports*. 12.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). *Why Hypothesis Testers Should Spend Less Time Testing Hypotheses*. 12.

Tipples, J., & Pecchinenda, A. (2019). A closer look at the size of the gaze-liking effect: A preregistered replication. *Cognition and Emotion*, *33*(3), 623–629. https://doi.org/10.1080/02699931.2018.1468732

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PLoS ONE*, *8*(3), e59202. https://doi.org/10.1371/journal.pone.0059202

Veer, A. E. van ât, Gallucci, M., Stel, M., & Beest, I. van. (2015). Unconscious deception detection measured by finger skin temperature and indirect veracity judgmentsâ"results of a registered report. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00672

Zgonnikov, A., Atiya, N., O'Hora, D., Rano, I., & Wong-Lin, K. (2019). *Beyond reach: Do symmetric changes in motor costs affect decision making? A registered report* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4bjeg

Zopf, R., Butko, M., Woolgar, A., Williams, M. A., & Rich, A. N. (2018). Representing the location of manipulable objects in shape-selective occipitotemporal cortex: Beyond retinotopic reference frames? *Cortex*, *106*, 132–150. https://doi.org/10.1016/j.cortex.2018.05.009