Università degli Studi di Padova

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di laurea magistrale in Psicologia Clinico-Dinamica

Tesi di laurea magistrale

Pianificazione della numerosità campionaria basata sul giudizio degli esperti e sintesi della letteratura. Una proposta per stimare dimensioni dell'effetto clinicamente rilevanti.

Sample size planning informed by experts and literature: toward clinically meaningful effect sizes.

Relatore:

Prof. Gianmarco Altoè

Laureando:

Emanuele Bollini

Matricola: 2119954

Correlatori:

Perugini Ambra, Sità Laura

Anno Accademico 2024/2025

Acknowledgements

Some really fancy acknowledgements here

Table of contents

Chapter 1

Replication crisis in psychology

In this chapter we outline key methodological and conceptual issues underlying the replication crisis, introducing the concept of replication and its role in science, then addressing major problems such as low replication rates, weak theories, poor measurement, and misleading statistical practices. After delimiting the inferential paradigm, we will then focus on the misuse of the inferential paradigm, especially the overreliance on Null Hypothesis Significance Testing (NHST).

1.1 Introducing the replication crisis

In recent years, psychology has faced a credibility crisis, raising doubts about the credibility of many published findings (Ioannidis, 2005). Since the beginning of such crisis, significant efforts have been devoted to improving data analysis, theory testing, and Open Practices (Lakens, 2019; Nosek et al., 2015; Nosek & Lakens, 2014; Oberauer & Lewandowsky, 2019).

We align with the interventionist perspective that identifies superficial methodological choices (specifically, the neglect of rigorous sample size planning and clinically meaningful effect size inter-

pretation) as one of the main causes of the crisis. In this thesis, we will develop and substantiate this explanation, arguing that a principled application of the Neyman-Pearson framework provides the necessary corrective to these methodological shortcomings.

1.2 What is a replication

"Scientific progress is a cumulative process of uncertainty reduction that can only succeed if science itself remains the greatest skeptic of its explanatory claims" (Abraham et al., n.d., p. 7).

Arguably, the very foundation of science depends on the replication of studies (Murphy et al., 2025). The greater our ability to replicate findings over time, the more confidence we gain in our beliefs (Collaboration, 2015; Schmidt, 2009). Therefore, replication allows us to be more skeptical about certain theories (Eronen & Bringmann, 2021), while also more confident in confirming and generalizing others (Cohen, 1994; Schmidt, 2009).

It is important to define terms that are often used interchangeably in the literature, despite referring to distinct concepts. A *reproduction* study consists in observing the same results, performing the same statistical analysis with the same data (indeed, it's also often referred to as "computational reproducibility"). In contrast, *replication* study consists in conducting the same study with different data and obtaining similar results (Nosek et al., 2022). The main challenge in replication studies is to determine whether the results we obtain are sufficiently close to the original ones, to consider the replication successfull.

Replication is often divided in two categories. Replication that aims to closely match the original study is called *direct replication.* Direct replication are very useful to confidently identify false positive (Nosek & Lakens, 2016). On the contrary, when variations are introduced in the study (e.g., different population), this is reffered to as *conceptual replication* (Derksen & Morawski,

2022). Conceptual replication may be useful to explore if the phenomenon can occur in different scenarios, thereby enhacing our comprehension of the effect, and offering more support to the theory (Nosek & Lakens, 2016). However, some argue that conceptual replications often attribute the effect to changes in study design rather than challenging the phenomenon itself (Nosek & Lakens, 2016). Offering a different perspective on the matter, Machery (2020) proposes abandoning the direct/conceptual distinction, and proposes instead to classify replications on how many components are resampled (e.g., units, treatments, measurements, settings): each combination of resampled components examines a different aspect of the original experiment's reliability.

1.3 Replication crisis

In recent years, many studies have shown low replication rates in psychological research. This has led to the conclusion that many psychological findings may actually be false (Collaboration, 2015; Ioannidis, 2005; Machery, 2020; Malich & Munafò, 2022; Oberauer & Lewandowsky, 2019; Vazire et al., 2022).

The main reason why positive results can be easily produced, and therefore published, is due to some inappropriate methodological practices: the most relevant among these are the misuse of an inferential procedure known as the Null Hypothesis Significance Test (NHST). and a range of questionable research practices (QMP). A larger overview regarding of these two phenomena will be proveided in the dedicated section (Head et al., 2015; Ioannidis, 2005; Scheel, 2022).

The spread of false positives is worsened by a publication bias. Publications offering significant results tend to be published more frequently (Scheel et al., 2021). There's a substantial pressure to publish, and studies presenting significant effects are often published at higher rate (Banks et al., 2016; Primbs et al., 2023). At the same time, likely due to the inherent difficulty in replicating

studies in psychology (Hall III, 2023), scientific publishing world prioritizes novelty over replication (Collaboration, 2015). Which contributes to the dissemination of low-quality or non-replicable findings.

This controversial finding undermines the very foundation of science, as replication is typically the primary mean by which we identify theories that are likely true. However, since replication is rarely done in psychology, overlapping theories that have not been decisively falsified yet remain common in the field (Eronen & Bringmann, 2021).

1.4 A brief mention to a conceptual crisis

The replication crisis has drawn attention to additional issues that characterize the field of psychology, many of which originate from the discipline's conceptual foundations and are worth noting: the validity crisis and the theory crisis. These aspects are broad and have been extensively discussed elsewhere. Here, a brief explanation will be provided, for the sake of completeness, before proceeding with the discussion of the replication crisis.

1.4.1 Validity Crisis

Validity crisis concerns the precise definition of the object of study within a given field (Eronen & Bringmann, 2021). If researchers are unable to define what they aim to measure, it will be unclear whether they will have been measuring the intended construct in the first place (Flake & Fried, 2020). In recent years, in response to the replication crisis, increasing attention has been given to the foundations of psychological measurement, as many studies fail to provide clear definitions of the constructs of interest (Vazire et al., 2022). Earlier, Flake and Fried (Flake & Fried, 2020) introduced the concept of *questionable measurement practices* (QMP), addressing all those practices

that cast doubt on the validity of measurement, such as lack of transparency in how measures were used and superficial or inappropriate application of measurement tools.

Following the recognition of this crisis, several recommendations have been proposed to improve measurement quality. For instance, Flake (Flake & Fried, 2020) emphasizes the need for greater transparency, and provides a structured series of questions to help researchers define construct and avoid QMP.

The crisis of measurement validity is also reflected in the weakness of psychological theories, which are often based on poorly defined constructs.

1.4.2 Theory crisis

Theory crisis refers to "a weak logical link between theories and their empirical test" (Oberauer & Lewandowsky, 2019, p. 2). This is arguably a more fundamental problem: theories are not strong enough to derive strong hypotheses or to predict the dimension of any effect (Eronen & Bringmann, 2021; Fried, 2020). Without strong hypotheses, theories cannot be falsified. Moreover weak theories can easily be defended after the results are known by adding auxiliary hypotheses. A common line of defense is the appeal to "hidden moderators", which are often invoked to explain failed replication (Fried, 2020). If that is the case though, we are not able to tell whether data corroborates our theory or not.

As Eronen and Bringmann highlighted we should invest more energy into building stronger and more precise theories (Eronen & Bringmann, 2021). However psychology is a challenging field, and to build strong theories we first need to identify clear phenomena. Therefore, some argue they should focus more on exploratory research (Fife & Rodgers, 2022; Oberauer & Lewandowsky, 2019), and in phenomenon-driven research, as it helps constrain the space of plausible theories (Eronen &

Bringmann, 2021).

While the validity crisis and the theory crisis certainly warrant further investigation to strengthen psychology's conceptual foundations, the replication crisis is more commonly attributed to other factors. The two main areas of concern are questionable research practices and misunderstandings related to inferential methodology

1.5 Questionable Research Practices

Questionable Research Practices (QRPs). QRPs are all the methods implemented in the field that aim to obtain a significant value, such as changing the hypothesis after the results are known (often referred to as "HARKing"), stopping data collection after achieving the desired result (often refferred to as "optional stopping"), collecting more data after seeing whether results were significant, manipulating statistical analysis in order to obtain significance (often referred to as "p-hacking"), presenting the results of a study that best support the hypothesis instead of reporting all the findings (oftentimes also called "cherry picking") and many others (John et al., 2012; Scheel et al., 2021). Therefore, QRPs include unethical and ambiguous practices, and have been shown to increase the likelihood of obtaining significant results (John et al., 2012), and increase the spread of false positive (Flake & Fried, 2020). Although often not driven by deliberate intent, QRPs are typically fueled by self-serving biases and the strong pressure to produce significant findings (John et al., 2012; Simmons et al., 2011).

To improve the situation, several solutions have been recently proposed, including strengthening statistical standards, conducting high-powered replications, providing open data, materials and algorithms to enhace reproducibility, and clearly distinguishing a priori hypothesis through preregistration (a practice aimed at preventing HARKing and clarifying the distinction between

confirmatory and exploratory research) (Nosek et al., 2015; Oberauer & Lewandowsky, 2019)

Significant results, however, refer primarily to outcomes derived from the most widely used statistical framework in psychology: null hypothesis significance testing (NHST) (Head et al., 2015). A brief clarification is useful here to understand its role in the replication crisis.

1.6 Inferential framework and its misunderstanding

To address the methodological issues at the core of the replication crisis, we will first specify the inferential framework we are adopting. Then, we will outline the Neyman-Pearson approach, as it is useful for understanding the weaknesses of the aforementioned NHST

1.6.1 Frequentist paradigm

As previously noted, we must first clarify the logic used to interpret empirical evidence. Among all the approaches used in inferential sciences, two in particular are mostly used, namely frequentist approach and bayesian approach. Although this thesis works within a frequentist framework, both approaches are sound; see Van de Schoot et al. (2014) for an for an overview of the Bayesian perspective.

In a frequentist perspective, probability is defined as the frequency (i.e., number of occurences) of an event in a set period of time (VandenBos, 2007). Having set the inferential paradigm of reference, we will now outline the problems that have emerged in the context of hypothesis testing, and which solutions have been proposed.

1.6.2 Neyman and Pearson approach

Within this inferential framework, several methodologies exist for hypothesis testing. One of the most studied is the Neyman-Pearson approach.

In the Neyman-Pearson approach, before collecting data, a predetermined level of inferential risk is established in order to make a decision between different hypotheses defined a priori. The test result will allow one to act as if one of the two hypotheses were true and the other false, given a certain level of risk. To do this, one must decide in advance on the following elements (Gigerenzer, 2004):

1) Defining two opposing hypothesis. The null hypothesis (H_0), which usually assumes no effect or difference, and the alternative hypothesis (H_1), which posits the presence of an effect and is the focus of testing. The magnitude of the effect has to be decided basing on theoretical or practical criteria. By defining H_0 and H_1, the associated sampling distributions are also specified. This allows for the division of the sample space into acceptance and rejection regions for each hypothesis.
2) To define the risks associated with the test. To do so, we specify in advance values for α (the probability of incorrectly rejecting the null hypothesis when it is true, namely Type I error) and β (the probability of failing to reject the null hypothesis when it's false). Power can be than calculated as the probability that the test has to reject the null hypothesis (H_0) when the alternative hypothesis ($H_{1)}$ is true (Altoè et al., 2020). The level of risk one intends to accept depends on the context and the phenomenon under investigation (Maier & Lakens, 2022).
3) Define the sample size. Once we know the a priori effect size, and we determine the values for α and β, the sample size is consequently determined in a design analysis.

The test then yields a p-value, which is the probability of observing the same or higher data than observed ones, if H0 is true. The p-value is then compared against the pre-specified α (therefore also called "significance level"), and we either accept H_1 and reject H_0, or reject H_1 and accept H_0 (Neyman, 1957).

The Neyman and Pearson actually originates from an other approach, the Fisher approach (Fisher, 1955). This approach only yields the p-value under H_0, so the smaller the p-value, the more the data are unlikely under the Null Hypohtesis. This approach though is intended for very preliminary analyses, and results must be interpreted basing on the context at hand (Gigerenzer, 2004). Specifying the usage and limitation of these procedures should make it easy to clarify the issues related to the previously noted NHST, which we will outline in the following section.

1.6.3 NHST approach

During its history, the development of psychology as a field has led to the widespread use of a simplified and less rigorous variation of this method, the aforementioned Null Hypothesis Significance Testing (NHST) (Gigerenzer, 2004). This test represents a hybrid form of the Neyman-Pearson approach and the earlier approach developed by Fisher (Fisher, 1955). The NHST paradigm works as follows (Gigerenzer, 2004):

- H_0 is defined as a null hypothesis (e.g., we test against the complete absence of an effect).
- No alternative hypothesis is specified.
- α is conventionally and uncritically set at 0.05.
- Finally, if p-value < 0.05 (statistically significant), the null hypothesis is rejected.

As previously noted, this methodology constitutes a simplified and weaker version of the Neyman-Pearson framework, as it omits two of its essential components. First, the significance level (α) is

commonly fixed at 5% without theoretical justification, making it an uncritical convention rather than a reasoned decision. In fact, it may be more important to identify false negatives more precisely in different context, e.g. a false positive that leads to a useless surgery may arguably be worse than uselessly having some therapy sessions, where we may accept higher risk for false positive. Second, since no alternative hypothesis is specified, no hypothesis can ultimately be accepted. This absence also prevents the definition of an ideal sample size in advance, often leading to under-powered studies (Gigerenzer, 2004). Moreover, without an alternative hypothesis, the a priori dimension of the effect cannot be calculated, making the interpretation more misleading.

To these characteristics, one must add a series of further elements of concern, as the NHST approach is characterized by a widespread misconception about logic of hypothesis testing and interpretation of p-values (Fife & Rodgers, 2022; Lakens, 2021). See Lakens (2016) for a more detailed discussion.

A vast majority of misconceptions consists in attributing to p-values informative strength over the theory to be tested (Gigerenzer, 2004). In fact, once the test yields a significant result, the research hypothesis, which had not been previously specified, is usually considered confirmed. This is incorrect, as no alternative theory had been previously specified. On the other hand, if a non-significant result is obtained, most psychologists often conclude that there is no effect, which is equally false (Lakens, 2017). Truly, a significant result does not indicate that a theory is likely true; it only tells us how likely the observed data are, assuming the null hypothesis is true (Cohen, 1994), since, as Gingerenzer states, "The probability $P(D|\mathrm{H}_0)$ is not the same as $P(\mathrm{H}_0|D)$, and more generally, a significance test does not provide a probability for a hypothesis" (2004, p.95). Ultimately, p-value is often erroneously considered as the strength of an effect or relationship (Head et al., 2015).

In synthesis, even when properly understood, NHST presents some unavoidable weaknesses. It only allows for the rejection of the null hypothesis, offering little to no meaningful information about the interpretation of the effect (Cohen, 1994). Moreover, since population parameters in the real world are virtually never exactly zero, even negligible differences can become statistically significant with sufficiently large sample sizes (Lin et al., 2013). This can lead to a misleading sense of confirmation for effects that are practically irrelevant (Lin et al., 2013; Malich & Munafò, 2022). Additionally, by omitting the alternative hypothesis and the corresponding effect size, NHST prevents the calculation of an appropriate sample size to achieve a desired level of statistical power (Gigerenzer, 2004).

1.7 Aim of this thesis

As shown in this chapter, improving inferential methodology and addressing the replication crisis requires a broad set of interventions. One intervention is strengthening inference by applying the Neyman-Pearson framework as originally intended. Doing so, this thesis directly addresses two critical aspects of robust inference, often neglected in recent times: rigorous sample size planning and meaningful effect sizes interpretation.

To do this effectively, it is essential to prespecify and formalize both the null and alternative hypotheses when designing a statistical test (as we outlined, when one is not able to formalize both hypothesis, more exploratory research should be done.) This approach avoids some of the major limitations of Null Hypothesis Significance Testing (NHST). As discussed, NHST does not support prospective power analysis, which is necessary to determine an adequate sample size. As a result, many studies are underpowered. Moreover, statistical significance is sometimes achieved for trivially small effects, leading to results that are only weakly informative. Finally, researchers

frequently misinterpret p-values as indicators of the strength of an effect.

As we have shown, many of these issues can be addressed by explicitly considering the effect size under the alternative hypothesis. This allows for proper planning of adequately powered studies and requires a theoretical estimation of the expected effect, thereby emphasizing the importance of effect sizes.

In chapter two, we will demonstrate that interpreting effect sizes is crucial for obtaining meaningful results, although they are often assessed using uninformative benchmarks. We will introduce two approaches to ensure more meaningful consideration of effect sizes. The first is the "plausible effect size" (Gelman & Carlin, 2014), which helps define an effect size based on what is realistically expected in the context of the specific test. The second is the "Smallest Effect Size of Interest" (Anvari & Lakens, 2021), also known as SESOI, a concept recently applied in psychology that can help identify practically relevant effects. Finally, we will propose a method to compare these two approaches, assessing whether the effect sizes we consider meaningful are likely to be detected in practice.

Having established the foundation of this thesis, in chapter three we will explore expert elicitation procedures—methodologies for deriving effect sizes based on expert judgment—which we will use to determine the SESOI for the present research.

In Chapter Four [...]

Chapter 2

Enhance effect size interpretation

"Thinking hard about effect sizes is important for any school of statistical inferences [i.e., Frequentist or Bayesian], but sadly a process often neglected."

— Dienes, 2008, p.92

In this chapter we will examine the effect size, providing a definition and a brief explanatory example. We will then outline its interpretation challenges, adressing the limits of statistical significance compared to the concept of practical significance. The distinction will allow us to introduce methods for formulating meaningful hypotheses, therefore explaining the plausible effect size and the smallest effect size of interest (SESOI) to enhance inferential robustness. Then we will propose a practical methodology as a combination of the two above.

2.1 Why we need effect sizes

In the previous chapter, we highlighted the importance of formalizing an alternative hypothesis while doing hypothesis testing within the Neyman-Pearson framework, or, in other words, the importance of specifying an effect size for the study. As outlined, this would improve research in the social sciences in several ways. Firstly, by allowing us to condutct a proper design analysis, it would enable transparent preplanning of an appropriate sample size, thus reducing the risk of false positives (Altoè et al., 2020). Secondly, the ability to predict effects and then observe whether they are confirmed or disconfirmed would lead to more falsifiable studies (Anvari & Lakens, 2021). Furthermore, since p-values alone can lead to the acceptance of results with low practical significance, incorporating effect sizes reduces the risk of attributing importance to potentially meaningless findings. These advantages highilight why effect sizes are essential complements to significance testing in study design and in result interpretation.

2.2 Effect size

As Peck and Flora report (2018, p.209), "the term *effect size* literally translates to some magnitude (or size) of the impact (or the effect) of a predictor or an outcome variable". Briefly said, it quantifies a given effect. As the definition depends on the phenomenon of interest, there is a very wide amount of possible definition to determine an effect size (Kelley & Preacher, 2012). In psychology research, though, effect sizes are most often used to indicate a relationship or a difference between two variables (Borenstein et al., 2021).

There are two main ways to report effect sizes in the literature. Firstly, the raw mean difference,where the effect size is expressed in the same units as the original measurement scale

(Borenstein et al., 2021). For example, the raw mean difference could quantify the mean difference between two groups on blood pressure measurement. Raw effect sizes offer a direct and intuitive idea of the meaning of the effect, especially given a widespread unit scale (as mmHg for the given example)(Borenstein et al., 2021). Nevertheless, many researches in social sciences use a variety of different measurements, adotping different scales, making it challenging to have a shared understanding of raw effect sizes (Altoè et al., 2020).

Given this limitation, researchers often prefer a second approach, i.e. standardized effect sizes, which facilitate comparison across studies and contexts. For this purpose, many different standardized effect sizes exist. In addressing this matter in the present thesis, however, we will focus on the most widely used and well-known measure (Altoè et al., 2020), namely Cohen's d (μ_{d}) (Cohen, 1988). In a concise formulation, Cohen's d is the raw difference between two popoluation means (μ_{A} μ_{B}) devided by the common standard deviation (σ):

$$\delta = \frac{\mu_A - \mu_B}{\sigma}$$

This formulation allows for standard interpretation across different measurement scales, which enhances comparability in psychological research.

2.2.1 Illustrating effect size

To illustrate how a standardized effect size such as Cohen's d operates in practice, we now present a simplified example based on a pre-post intervention study in a clinical context. In such studies, psychological distress is typically measured before and after an intervention using standardized instruments, for example, the Kessler Psychological Distress Scale (K-10). Participants are randomly assigned to a treatment group or a control group, and the effect of the intervention is assessed by

comparing changes in K-10 scores. In the following example, intended solely for didactic purposes, we provide a visual and conceptual representation of what it means to specify an effect size under the null and alternative hypotheses. Specifically, our alternative hypothesis specifies an expected standardized effect of $d = 0.41$.

The figure 2.2.1 shows two theoretical distributions of the test statistic under the null hypotheses (H_0) and alternative hypotheses (H_1), within a Neyman-Pearson framework. These distributions derive from a comparison between a treatment group and a control group in a pre-post design. The black curve represents the distribution under the null hypothesis (H_0), which assumes that the intervention has no effect (i.e., the mean change in the treatment group is equal to that in the control group) corresponding to an effect size of $d = 0$. The blue curve represents the alternative hypothesis (H_1), assuming a true standardized effect size of $d = 0.41$, meaning the treatment group improves, on average, by 0.41 standard deviations more than the control group. The distance between the two distributions reflects the assumed effect magnitude.

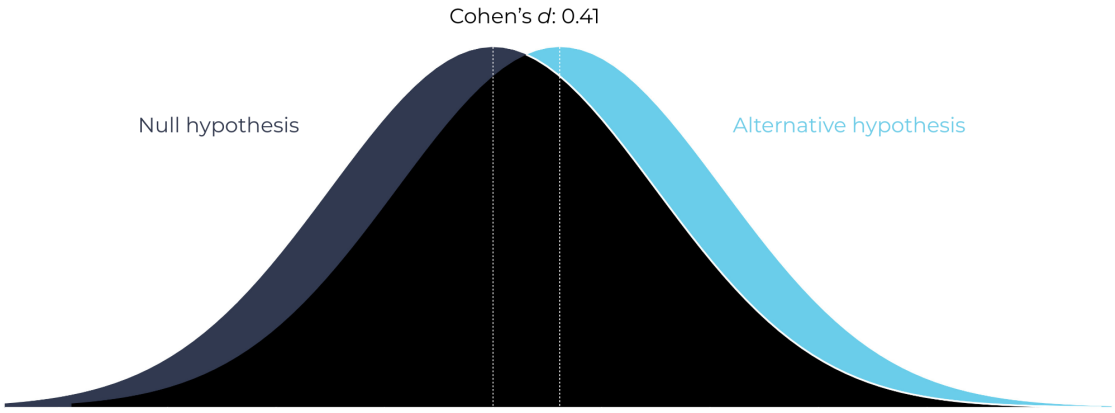

Figure 2.2.1: A graphical rapresentation of a Cohen's d, in a RCT group. (Magnusson, 2023)

We have provided a graphic illustration of an effect size, which represents values distribution when the treatment has no effect (H_0) and when the effect is 0.41 (H_1). It is quite intuitive

though that neither a graphical representation nor a numerical value clearly informs us of *how much* of improvement patient have experienced under the alternative hypothesis. This example introduces the main matter discussed in this thesis that we will detail in the following section, that is interpretation of effect size.

2.3 Interpretation of effect sizes

When addressing effect sizes, it is important to distinguish between informative and uninformative ones. While this may not be necessary in all sciences, it goes without saying that in psychology we must be able to differentiate the magnitude of effects. Taking into account the clinical context, one in which this distinction is especially critical, it is clear that we need to discriminate between mild and highly effective interventions, just as we distinguish between severe and modest clinical cases.

However, in psychology, there is a tendency to treat any effect size as a success (Hilgard, 2021), which hinders a clear distinction between these two categories. To counter this issue, it is useful in this regard introducing the distinction between practical and statistical significance, which are not alternatives, but complementary dimensions of result interpretation.

2.3.1 Statistical significance and practical significance

Statistical significance is met when a certain p-value indicates that the result meets the level of evidence required by the researcher's chosen threshold (α) to reject the null hypothesis (H_0), but p-value alone provides no information about the magnitude or practical relevance of the observed effect (Kelley & Preacher, 2012; Pek & Flora, 2018). A statistically significant result may correspond to a trivially small effect, especially in studies with large sample sizes, while relevant effects may go undetected in underpowered studies. Practical significance, instead, refers to the distinction

between interesting and uninteresting effects (Anvari & Lakens, 2021). In fact, as already shown, effect sizes themselves are not automatically informative of anything, and should be interpreted in light of substantive criteria (Grissom & Kim, 2012; Primbs et al., 2023).

2.3.2 The misuse of Cohen's classification benchmarks

These points illustrate the necessity of evaluating effect sizes within the specific context and goals of research. Despite this, the interpretation of effect sizes is usually related to a classification proposed by Cohen (Cohen, 1988), who suggested values of $d = 0.2$, $d = 0.5$, $d = 0.8$ as indicative of "small", "medium" and "large" effect sizes respectively.

A graphical rapresentation of the three effect sizes is reported at figure 2.2.1, 2.2.2 and 2.2.3 respectively. These sets of distributions derive from a hypothetical comparison between a treatment group and a control group in a pre-post design. The black curves always represent the distribution under the null hypothesis (H_0), which assumes that the intervention has no effect (i.e., the mean change in the treatment group is equal to that in the control group) corresponding to an effect size of $d = 0$. The blue curves always represent the alternative hypothesis (H_1), assuming a true standardized effect size of of $d = 0.2$, $d = 0.5$, $d = 0.8$, meaning the treatment group improves, on average, by 0.2, 0.5 and 0.8 standard deviations more than the control group, respectively. The distance between each pair of distributions reflects the assumed effect magnitudes.

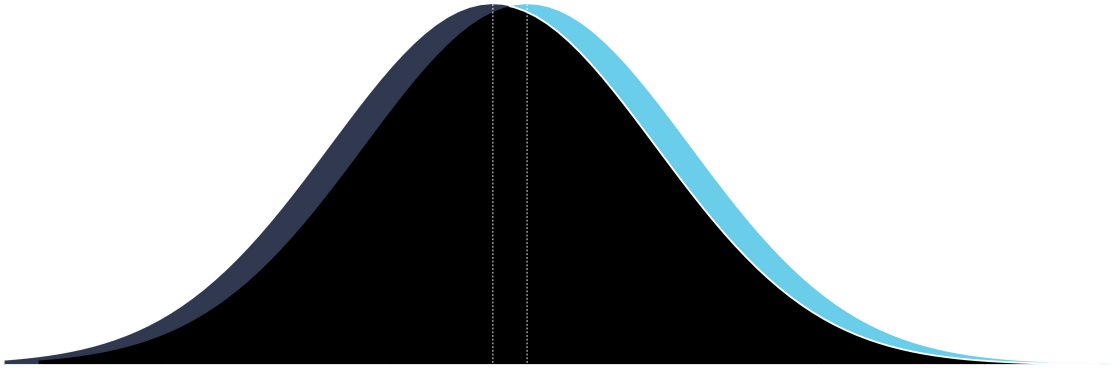

Figure 2.2.1: A graphical rapresentation of 0.2 Cohen's d (Magnusson, 2023).

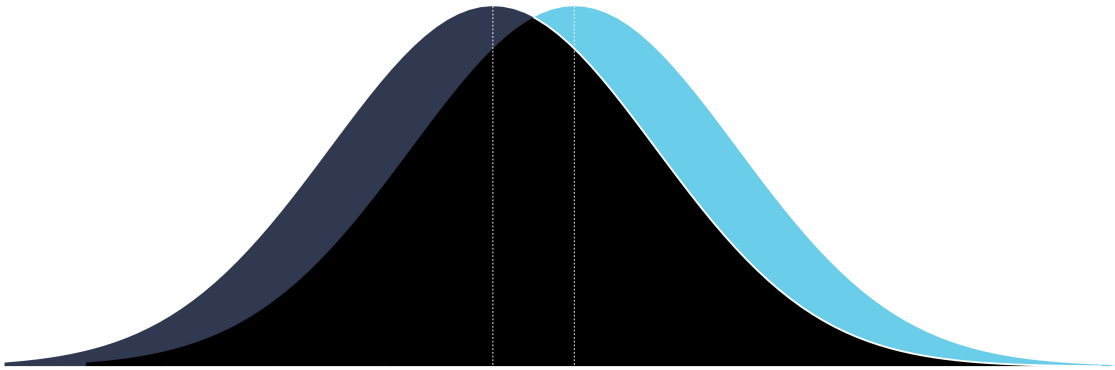

Figure 2.2.2: A graphical rapresentation of 0.5 Cohen's d (Magnusson, 2023).

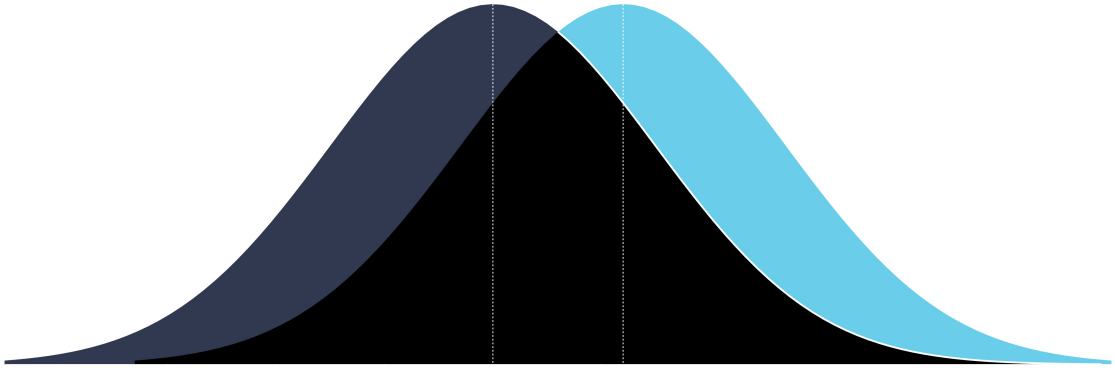

Figure 2.2.3: A graphical rapresentation of 0.8 Cohen's d (Magnusson, 2023).

While Cohen himself intended these benchmarks as a last resource reference, psychology research has adopted them as standards. Unfortunately, applying these benchmarks without considering the context leads to uninformative conclusions (Funder & Ozer, 2019). Effect sizes require different interpretations depending on the field of interest, the specific content, and the research methods used, as in a given field a small effect size can be extremely relevant in practice(Altoè et al., 2020).

2.4 Alternative approaches to interpret effect sizes

This clarification has motivated the development of more appropriate benchmarks for effect sizes that are more context specific and informative.

A first category of solution proposed by Funder and colleagues (2020), is comparing findings to well-established results in the literature, e.g. one could compare effect sizes typically considered "small" with classical results obtained in studies on attitude change (Festinger & Carlsmith, 1959), or with to others non-psychology-related phenomena. However, by itself, this approach does not fully resolve the issue. Even if it could be useful to know whether results are comparable to those of

another study, if there is no concrete or meaningful sense of the effect in the previous study, we know merely that the effects are similar, but nothing about the meaning they convey. The implications are even less clear when applied across different fields.Therefore, solutions relying solely on conventional or representative benchmarks are unlikely to be sufficient.

To establish more voncingin benchmarks though, several methodologies have been proposed. Two of the main approaches are defining a "plausible effect size" (Gelman & Carlin, 2014) and identifying the "smallest effect size of interest" (SESOI) (Anvari & Lakens, 2021), both of which will be outlined below.

2.4.1 The plausible effect size

"The plausible effect size refers to what could be approximately the true value of the parameter in the population" (Altoè et al., 2020, p. 5) This concept was introduced to address the challenges of power analysis—or more broadly, design analysis (Gelman & Carlin, 2014), by helping researchers pre-specify an alternative hypothesis, thereby enhancing inferential strength for the reasons previously discussed.

Although the validity of the inferential process is generally strengthened when hypotheses are derived from theory and formalized in statistical terms, precise hypotheses are often not yet feasible in psychology for the reasons previously discussed. Therefore, their formalization can instead be based on literature reviews and/or meta-analyses (Altoè et al., 2020).

The main advantage of plausible effect sizes is to enable a critical interpretation of the results, effectively discriminating plausible from implausible results. With this methodology, we can avoid adjusting effect sizes, either implicitly or explicitly, to justify the value of a given sample size, as it's often done (Gelman & Carlin, 2014). As stated by Altoè "in general when observed effect size falls

outside the pre-specified plausible interval, we can conclude that the observed study is not coherent with our theoretical expectations. On the other hand, we could aslo consider that our plausible interval may be unrealistic and/or poorly formalized." (2020, p.9)

Although this metodology is useful to establish criteria that make us cautious about implausible effects, its direct impact on the inferential process is limited, as they do not directly provide with guidelines to distinguish which effect sizes should be deemed meaningful within a given line of research. In exploring further alternative solutions to this issue, the second aforementioned approach will now be explained.

2.4.2 The smallest effect size of interest

In the context of clinical research, scientists are particularly interested in the practical significance of a certain quantity of change. The first relevant ideas on this matter have emerged from the medical field, where quantifying the effect of a given phenomenon is clearly essential. These methods aim to quantify a *Minimal Important Difference* (MID). Subspecifications of this concept include the *Minimally Detectable Difference* (MDD), which emphasizes the ease of detecting a given difference, and the *Clinically Important Difference* (CID), which is based on relevant clinical outcomes such as recurrence or risk of rehospitalization (Norman et al., 2003).

One similiar concept has been recently proposed in psychology, the aforementioned *Smallest Effect Size Of Interest* (SESOI). The SESOI can be defined as "the smallest change that is needed in the outcome measure for people to subjectively notice and report a difference in how they feel" (Anvari & Lakens, 2021, p. 1). This concept translates the notion of an "important" change, defined as one that is both minimal and noticeable, into psychological research, providing a practical criterion for assessing the relevance of effect sizes.

The idea is to assess the importance of a given amount of change based on an external practical criterion. As preivously discussed, such a criterion cannot yet rely on solid theoretical reasoning (Riesthuis et al., 2024). However, several methodologies have been proposed, though still being at an exploratory stage. The most prominent among these include cost-benefit analyses, anchor-based methods, and consensus methods (Anvari & Lakens, 2021).

1) Cost–benefit analysis evaluates whether the observed improvement justifies the cost of the intervention, usually in comparison with alternatives. However, as noted by Anvari and Lakens themselves, "in basic psychology research, costs and benefit are not easily quantified" (2021, p.2).

2) Anchor based methods use a retrospective judgment as a reference to determine whether participants have improved, stayed the same, or worsened over some period of time (Anvari & Lakens, 2021; Lydick & Epstein, 1993; Norman et al., 2003). For example to estimate how much change has been experienced after a treatment, participants complete the measurement of interest before and after the treatment. They are then asked to indicate how much change they notice. Responses may vary depending on the rating scale used. These ratings can be collected through self-reports or clinician's assessments, and are called clinical "anchor". Then, the amount of change (effect size) corresponding to each category is measured (Anvari & Lakens, 2021). However, these methods are quiet recent, and not very much established.

3) Consensus methods are a recently proposed solution that involve asking experts for their opinion on what could constitute the smallest effect size of interest. Researchers then assess whether there is general agreement on the SESOI (Anvari & Lakens, 2021; Riesthuis et al., 2022).

Together, these methods provide preliminary yet important frameworks for defining practically meaningful effect sizes in psychological research.

2.5 A methodology for building a plausible SESOI

Based on what has been explained so far, it should be clear how a well-established and pre-specified effect size can strengthen inferential process and enhance the overall credibility of psychological research. In particular, the concept of a plausible effect size provides a solid basis for defining reasonable intervals for the true effect of a phenomenon. Conversely, methods for determining a Smallest Effect Size of Interest (SESOI) offer valuable guidance for identifying effects that are practically meaningful.

DOVREMMO DIRLO DA QUALCHE PARTE CHE I SESOI DA NOI CALCOLATI SONO EPISTEMICAMENTE DIVERSI DAI PLAUSIBLE

In the following table, we provide a concise summary of the advantages and disadvantages of both approaches.

	Plausible effect size	**SESOI**
Pros	• Evidence based	• Practically informative
Cons	• Practically uninfromative • Requires high-quality meta-analysis	• Strongly relies on subjectivity • Requires attentive esteems

Building on this foundation, the aim of this thesis is to propose a methodology that uses both concepts in a complementary way. Specifically, we seek to assess whether the SESOI identified is contained within the plausible interval for the effect size or, conversely, whether the effect sizes considered plausible include effects that are practically meaningful, thus ensuring that the estimated

true effect is not only statistically valid but also relevant.

The procedure goes as follows:

1) Since systematic reviews and meta-analyses can provide guidance on typical effect sizes (Gelman & Carlin, 2014), we can derive a plausible effect size using meta-analytic data. For illustration purpose, we assumed hypothetical meta-analytic results, evaluating efficacy of a clinical intervention, measured with a questionnaire that asses well-being levels. In this scenario, the meta-analysis indicates that the Coehn's d compared to the control was 0.44, with a 95% confidence interval from 0.27 to 0.61, indicating an improvement in well-being levels. We decide to set our plausible effect size within the confidence interval. A graphical hypothetical representation is shown in figure 2.3.1.

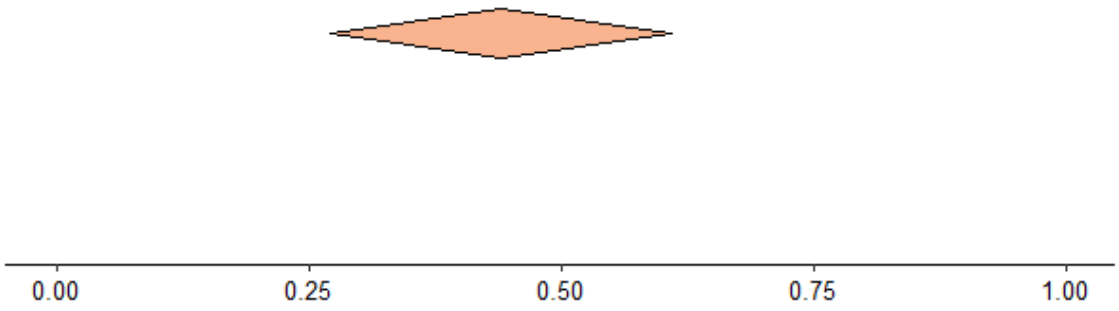

Figure 2.3.1: A graphical rapresentation of a hypothetical plausible effect size

2) As consensus methods are appropriate for guiding decisions in setting a SESOI that reflects practical relevance (Riesthuis et al., 2024), we will use this approach to generate one. We will give particular attention to the matter of consensus methods in the following chapter.

For clarity, we generated three different possible SESOI scores that may emerge from three different elicitation processes, as graphically represented in figure 2.3.2.

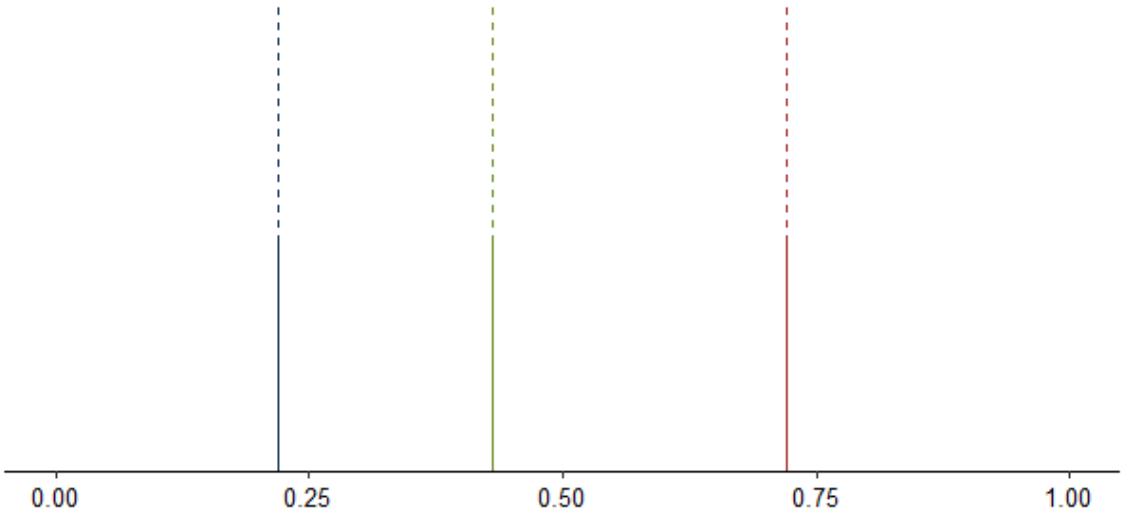

Figure 2.3.2: A graphical rapresentation of the hypothethical SESOIs

3) The two results are then compared, as illustrated in figure 2.3.3 and discussed. In the case of the first and smallest generated SESOI score, for example, we might think that since the plausible effect size falls after the line of clinical significance, the vast majority of results in this particular field of research are clinically relevant. With the second and middle-ranged SESOI score, we could say that most of the results in this field are somewhat clinically relevant. With the last and largest SESOI score, we might argue that our clinical interventions are clinically irrelevant, even though the effect size might sometimes be even medium or large, based on Cohen's standards. Connversly, we could also say that our questionnaire are not suited to effectively identify a clinically relevant change. Obviously, results are never so easily discussed, and deeper reasoning is yet to be done.

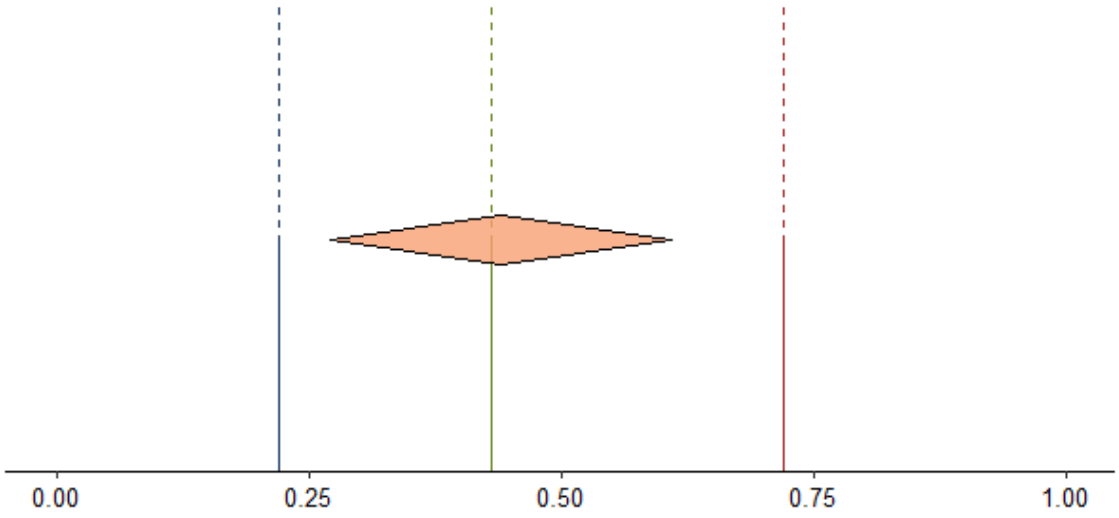

Figure 2.3.3: A graphical rapresentation of the comparison between the plausile effect size and the SESOI.

4) Finally, a design analysis will be conducted on the obtained result, to illustrate the practical utility of the proposed methodology.

As consensus methods have only recently and infrequently been used to assess a SESOI, the following chapter will focus on these approaches, and more specifically, on their more systematic form: expert elicitation.

2.6 Conclusions

In the previous chapter, we discussed how the NHST framework often leads researchers to overlook power, sample size, and the practical meaning of results, with p-values frequently misused as indicators of effect strength. Here, we focus on effect sizes as essential tools for addressing these issues, as they connect directly to both study planning and result interpretation.

We defined the effect size, outlined common interpretative pitfalls, and distinguished statistical from practical significance. To enhance interpretation, we presented two complementary approaches: the plausible effect size, based on prior evidence or theory, and the Smallest Effect Size of Interest (SESOI), based on practical relevance. We then proposed a method to integrate both, comparing whether plausible effects are also meaningful.

This chapter shows that careful consideration of effect sizes strengthens inference and improves research utility. In the next chapter, we will explore expert elicitation as a method to define the SESOI within our framework.

Chapter 3

Expert Elicitation

In this chapter we will examine expert elicitation, providing a definition and a brief explanation. We will then outline its main challenges, addressing issues related to expertise, estimate quality, cognitive biases, and aggregation methods. This distinction will allow us to introduce the leading protocols for expert elicitation, explaining how structured approaches can improve accuracy and decision-making. We will then discuss the potential application of these methods for establishing a SESOI. Finally, we will introduce a tailored implementation of a remote elicitation.

3.1 Why we need expert elicitation

In the previous chapter, we focused on effect size interpretation as a crucial aspect for addressing sample pre-planning and H_1 specification. Having distinguished between statistical and practical significance, we outlined how the numerical value of the effect size alone does not inform us directly about any theoretically nor practically meaningful implication. To enhance interpretation, we then proposed a method to integrate the plausible effect size and the Smallest Effect Size of Interest

(SESOI). As explained, the two are complementary, as the plausible effect size informs about the statistical plausibility of our findings while the SESOI could give us a precious practical insight.

As we have already outlined, SESOI can be obtained based on different methods, i.e. cost-benefit analyses, anchor based methods and consensus methods. Having already discussed the limitations of the first two, we will the outline the advantages of consensus methods.

Firstly, by relying on consensus methods we may ensure the obtained effect size is linked to its practical or clinical relevance. In practice, clinical relevance is often determined by expert clinicians' judgement, where clinicians must take critical decisions. We argue that relying on experts' judgement has been our reliable approach so far, and could represent a solid ground on which to establish our SESOI. This is especially true as long as we find a way to systematize said opinions, and until more precise and accurate criteria are established. Secondly, consensus methods methodologies are well known and extensive literature has already been produced on this subject, making its implementation both feasible and sound.

While at a first glance it may be questionable to rely on experts' judgement, in fields where a value is not known, or decision making is crucial, it is reasonable to rely on experts' opinions (Authority, 2014b). This is even more reasonable since one's subjective opinion can be collected in a well established procedure, which is mostly known as Expert Knowledge Elicitation (EKE) (O'Hagan, 2019).

In this chapter we will provide an explanation of such methodologies. We will then introduce the protocol we will be using to collect the SESOI in chapter four.

3.2 What is expert elicitation

Expert Knowledge Elicitation (EKE) is a well-established and sound procedure that incorporates expert judgement into formal analyses. The literature on EKE is extensive and covers many fields, including statistics, management sciences, economics, and environmental sciences (O'Hagan, 2019).

Generally speaking, *elicitation* is the process of gathering information needed for a given reason (Authority, 2014a). In EKE, this process refers specifically to obtaining information from one or more experts (Authority, 2014a; O'Hagan, 2019), and in research context, elicitation is usually used to obtain and formalize expert knowledge in the form of probability distributions or estimates (O'Hagan, 2019).

In this section, we will address four interconnected issues in EKE: the matter of expertise, the quality of estimates, cognitive biases, and the aggregation problem. First, the definition and selection of expertise determine the validity of the elicitation process, as only well-qualified experts can provide meaningful judgements. Second, the notion of a "good" estimate is examined through different criteria (accuracy, bias, calibration, and reliability). Third, since research has shown that expert opinion is not immune to systematic distortions, we will adress the matter of biases. Finally, the aggregation problem is discussed, as expert elicitation usually involves multiple contributors but ultimately requires a single output for decision-making.

3.2.1 The matter of expertise

The definition of "expert" and "expertise" is a key issue in EKE. Various definitions exist, depending on the purpose, methodology, and underlying theories (Authority, 2014a; Burgman, 2015; O'Hagan et al., 2006), but there is no clear or universally accepted distinction between lay and expert judgment (Burgman, 2015). Even though the debate on this matter is ongoing, when running a

study it is essential to clearly specify the type of expertise required and the criteria for expert selection (Authority, 2014a). Linked to this issue, an other significant issue revolves around the number of experts involved in the elicitation process. From a theoretical perspective, involving as many experts as possible is advantageous, as it improves the accuracy of estimates for the variables of interest. However, empirical evidence suggests that including too many experts becomes problematic. The quantity clearly depends on the field and study, but generally speaking, it is recommended to include a minimum of 5 experts, and no more than 8 to 15. Using too few will limit the diversity of perspectives and the informativeness of the results, while involving too many will add little value to the final estimates, while excessively increasing cost and time requirements (Authority, 2014a).

3.2.2 The matter of estimates

It is common practice, and almost common sense, to seek expert opinion when facing uncertainty or lack of knowledge on a particular issue. EKE formalizes this reasonable behavior by asking experts estimates for a given quantity, therefore we shall wonder what defines a "good" estimate. There are several criteria which are used to evaluate expert estimates, but the most typically used are accuracy, bias, calibration, and reliability (Burgman, 2015) :

- *Accuracy* is how close the expert's quantitative estimate is to the true value.
- *Bias* measures a consistent deviation from the true value, in one direction.
- *Calibration* is a measure of how often the expert's esteemed intervals contain the true value.
- *Reliability* addresses the characteristic of the expert, as it reflects the degree to which their estimates are repeatable and stable over time.

Not all of these factors are equally important, and the balance depends on the specific problem

at hand, as in some circumstances some criteria could be more relevant than others (Burgman, 2015). Although there is an open discussion about how accurate and reliable expert judgements are compared to non-expert opinions, evidence shows that experts generally provide better estimates within their area of expertise. For instance, reliability in EKE is often criticized, as judgements are not very stable over time, which reflects some of the compromises one must accept when using expert elicitation (Burgman, 2015). Nevertheless, we rely on expert advice when making decisions in situations where we lack sufficient information, and when it is our best or only source available (Burgman, 2015). For instance, in clinical psychology, where clinicians certainly benefit from using standardised measures and techniques, ultimately, decisions about assessment and intervention methods are based on the clinician's judgement. Therefore, considering these methodologies is also necessary to avoid naïve conclusions in clinical research.

3.2.3 The aggregation problem

As we already mentioned, in most cases, judgements are collected from multiple experts, as it increases the quality of the estimated value, but we typically require a single value as the final output for decision-making. This challenge is referred to as the aggregation problem (O'Hagan, 2019).

To address the aggregation problem, two main strategies are commonly employed, the "mathematical aggregation" and the "behavioural aggregation" (O'Hagan, 2019). The first solution is "mathematical aggregation", which involves eliciting individual judgements from each expert and fitting a probability distribution to each of them. These distributions are then combined into a single aggregated distribution by a mathematical formula. Such formula is known as a pooling rule. Several pooling rules exist, and choosing the most appropriate one is the most significant

choice when implementing this approach (O'Hagan, 2019). The second solution is the behavioural aggregation, which relies on interaction among experts. Firstly, experts discuss their views and reach a consensus. Then, a distribution is fitted to the group's collective judgment. This method, however, is so sensitive to interpersonal dynamics that may introduce bias. One example could be that dominant personalities can influence group discussions, even unintentionally, compromising the final result (O'Hagan, 2019). One of the most significant risks in behavioural aggregation was first noticed by Janis (1972), as the phenomenon of "groupthink". In this phenomenon, the desire for consensus overrides critical evaluation, leading to sub-optimal decisions.

3.2.4 The matter of biases

Such a reliance on experts' judgement makes it crucial to understand and consider potential biases that may influence expert decisions. Unstructured or naïve questioning can introduce unwanted cognitive biases, which can affect experts' judegements. Psychological research has identified ways in which superficial or unstandardised questioning can induce cognitive biases in expert judgment (O'Hagan, 2019). Below, we outline some of the most significant ones to provide an overview of the attention this issue has received in the literature:

- **Anchoring:** When asked to provide a numerical estimate, individuals tend to rely heavily on the initial value they consider (the "anchor"), and succesive adjustments remain biased toward that point (O'Hagan, 2019).
- **Availability:** Events that are more easily recalled or more memorable are often judged more probable, which can lead to an overestimation of dramatic or recent occurrences (O'Hagan, 2019).
- **Range-Frequency:** When the possible values of an uncertain quantity are divided into cate-

gories, experts tend to distribute probabilities quite evenly across those categories, regardless of the reasonable likelihoods (O'Hagan, 2019).

- **Overconfidence:** Experts often display overconfidence, which may result in confidence intervals that don't include the true value, therefore leading to low calibration values. This could be due to social pressures to demonstrate expertise, or due to the nature of elicited questions, which often involve less routine or familiar quantities. Sadly, habitual heuristics that serve experts well in everyday tasks may not be as effective in these elicitation contexts (O'Hagan, 2019).

To reduce the influence of these biases, questions in EKE are carefully designed and preplanned to minimise their impact (Authority, 2014a). To make expert knowledge as objectively as possible, elicitation must be carefully structured. This has led to the development of formal protocols, that are essential to enhance reproducibility and transparency while reducing the influence of biases. We will discuss said protocols in the following section.

3.3 The leading protocols

The decision to adopt either mathematical or behavioural aggregation methods has a direct impact on the conclusione we derive from expert judgements. Moreover, it gives even more importance to structuring proper protocols, which aim to reduce cognitive biases and enhance transparency and reproducibility throughout the process. These protocols differ greatly in their level of formalization, their approach to uncertainty, how they structure expert interaction and the aggregation process itself. Some of these have emerged as the leading ones in the literature. However, it is not possible to determine definitively which protocol is "the best" in terms of accurately capturing experts' knowledge and beliefs in the form of probability distributions. Even in the rare scenario where

the true values of the quantity of interest are eventually revealed, drawing reliable comparisons would require a substantial number of experts, randomly assigned to different protocols, and a large enough number of elicitation tasks across multiple scenarios for appropriate replication and generalization. Nonetheless, the matter of improving expert knowledge elicitiaion needs to be improved by innovative and systematic research efforts (O'Hagan, 2019).

EFSA recommends the use of three main elicitation protocols (Authority, 2014a).

1) The **Cooke protocol** is based on mathematical aggregation. It follows a structured approach known as the classical model. In this method, before providing estimates for the quantity of interest, experts first provide independent judgments on a set of *seed variables*. These seed variables are ulterior quantities to be asked, related to the nature of the target variables, but whose true value is known. The expert's accuracy in judging the seed variables serves as an indicator of the quality of their estimates for the unknown quantities, as this protocol assigns weights to experts based on their performance on the seed questions, and with these weights individual judgments are combined into a single aggregated distribution. The more precise one's seed estimate, the heavier its weight in the mathematical aggregation (O'Hagan, 2019).
2) The **Sheffield protocol** utilizes behavioural aggregation. It involves two distinct rounds of expert elicitation. During the first round, experts provide their individual assessments privately to a facilitator. These judgments are then shared and discussed collectively, with the goal of understanding the reasons behind the differences in opinion. Following this discussion, in a second round the group works together to reach a consensus judgment. This consensus is intended to represent the viewpoint of a rational, impartial observer, who should be able to compare objectively the different opinions. The protocol requires a skilled facilitator to guide the process, to ensure to minimize the disturbance of group dynamics and mitigate the biases

of the discussions (O'Hagan, 2019).

3) The **classic Delphi method**, finally, is mostly used to elicit judgments of uncertainty rather than simple point estimates. This approach combines elements of both mathematical and behavioral aggregation. Experts provide their judgments over two or more rounds, with feedback between rounds, to finally generate a collective response. Since anonymity of individual experts is crucial in this approach, there is limited interaction between experts, allowing some sharing of knowledge, while minimizing social biases risk. After the final round, a pooling rule is applied to combine the experts' distributions into a single aggregated judgment (O'Hagan, 2019).

Although these methods typically require a significant investment of time, effort, and financial resources, more agile and feasible versions have been developed. These include methods like the IDEA protocol (Hemming et al., 2018), which streamlines expert elicitation through two estimation rounds with intermediate discussion, enabling flexible remote implementation while maintaining methodological rigor.

3.4 Why eliciting a SESOI

As already discussed, elicitation protocols sure have many limitation and issues that must be taken into account. Nonetheless, where little information is available and decisions must be made, relying on the most accurate possible judgement we can elicitate from expert could help improve both research and practice within a given field.

Guidelines often recommend eliciting values rather than judgments (Hemming et al., 2018). But precisely because of the absence of clear guidance for clinicians, it's arguably even more crucial to establish a shared threshold through a trasparent, structured, and deliberative process that

minimizes individual biases. The very lack of objectivity of the construct at hand does not weaken the case for expert elicitation, but strengthens it further.

Moreover, we believe that expert elicitation could prove particularly valuable in the interpretation of effect sizes in psychological research, especially given the absence of clear criteria from either theory or other sources, as we outlined in chapter two and three.

3.5 A tailored procedure for eliciting SESOI

Since formal methods for a proper elicitation, such as the IDEA protocol, were not feasible within the constraints of this study regarding time and resources, a pragmatic, direct-to-expert approach was developed. This tailored procedure was designed to quickly capture the collective clinical intuition of practicing therapists, translating their experiential knowledge into a quantifiable benchmark for a minimal, yet meaningful, improvement.

In the following chapter, we will provide a more in depth explanation of its component and application to address the central aim of this thesis: to illustrate how structured expert judgment can aid in the interpretation of effect sizes and inform decision-making when theoretical guidance is lacking.

Chapter 4

Tailored elicitation procedure

In this chapter we will outline the methodology employed to define the smallest effect size of interest (SESOI) for the Kessler-10 (K-10) scale through a tailored elicitation process. We will address the issues of eliciting a SESOI, then we will give a brief introduction of the K-10. We will then explain the implemented procedure, along with its limitations.

4.1 The Kessler-10

The Kessler-10 (K-10) is a brief, 10-item self-report questionnaire designed to measure non-specific psychological distress, with a focus on symptoms of depression and anxiety. Respondents rate how frequently they experienced each symptom in the last 30 days on a 5-point scale (ranging from 1 = "None of the time" to 5 = "All of the time"). The total score ranges from 10 to 50, where higher scores indicate greater levels of psychological distress. Its brevity and robust psychometric properties make it a widely used tool in both research and primary care settings for screening and monitoring.

4.2 The elicitation procedure

Time and resource constraints prevented us from conducting a manualized elicitation, therefore we conducted a survey via e-mail, with a one-shot question to answer. The choice of this non-manualized methodology is considered appropriate within the context of a master's thesis, intended for illustrative purposes only, and serves as a substitute for a proper elicitation process.

We will now outline how the survey was constructed, how experts were recruited and data collected.

4.2.1 Survey structure

The survey was centered around a detailed clinical vignette depicting a university student with emerging depressive symptoms. The vignette was designed to illustrate a scenario where, after an initial treatment, the patient exhibits a minimal, yet meaningful, functional improvement. Neither lenght of the treatment nor type of treatment was detailed. This choice may be discussed, but it's our opinion that given the very definition of SESOI (the smallest effect size of interest) we are not interested in the time needed to obtained a relevant improvement, nor we are interested in how the improvement is obtained (whether it may be due to a certain treatment, or even spontaneous). Conversly, we are only interested in the score difference, given the relevant change.

Following the vignette, the survey presents the pivotal question: "Given such a relevant change, what difference in the total K-10 score would you expect?" (the original and complete versione can be fount in the appendix)

The methodology was inspired by the final two stages of the IDEA protocol (notably, "estimate" and "aggregate"(Hemming et al., 2018):

- Experts were asked to provide their best-guess estimates independently and privately within

a one-week timeframe, avoiding group dynamics like groupthink.

- Following the core principle of the final IDEA phase, the collected responses were mathematically aggregated (using measures such as mean or median) to derive a group estimate.

To refine the clarity and feasibility of the elicitation instrument, a pilot test was conducted. A sample of 6 master's students in clinical psychology was recruited. After being provided with a description of the K-10 scale, they were sent the survey. Their feedback and responses were used to identify ambiguities and ensure clarity.

Since no sensitive data were collected, ethical committee approval was deemed unnecessary. Participants were assured that their individual responses would be presented exclusively in aggregated form.

4.2.2 Expert recruitment and data collection

The target population for the elicitation consisted of practicing clinical therapists. The primary eligibility criteria were direct clinical experience and self-reported familiarity with the use of the Kessler-10 scale in their practice. Beyond these criteria, participants were selected based on a combination of convenience and availability.

Upon agreeing to participate in the data collection, participants were sent the questionnaire directly via email, without a preliminary explanatory phase.

The survey was distributed via email on Wednesday, 22 October at 09:15. A reminder was sent on the scheduled deadline day at 09:15.

4.3 Results

4.3.1 Limitations and emergent insights

- By eliciting only point estimates rather than uncertainty intervals, this approach does not capture their confidence levels and may reinforce overconfidence bias.
- Dal punto di vista operativo, la natura non verificabile del costrutto preclude l'applicazione di tecniche come il "performance weighting" (come descritto nel capitolo precedente), che richiedono domande a risposta nota per calibrare il peso degli esperti. Ciò ha reso inevitabile il ricorso a un'aggregazione a pesi uguali, un compromesso metodologico ampiamente accettato in contesti simili, sebbene comporti il rischio di diluire il contributo degli esperti più accurati.

To our knowledge, there are no precedents in the literature for elicitations of this type. This meant we were unaware of the potential associated problems.

We found that one challenge for clinicians is shifting their focus from "improvement given the treatment" to "score variation given the improvement."

Difficulty in finding clinicians with specific expertise in using a particular scale.

A key challenge was the inherent circularity in defining the "minimally important change" for expert elicitation. To operationalize the concept, we used a clinical vignette depicting minimal but meaningful functional improvement and asked experts to quantify the expected Kessler-10 score change. While necessary, this approach relied on clinicians' individual interpretation, potentially introducing more variability. This underscores the importance of eliciting multiple expert judgments to capture diverse clinical perspectives.

4.4 Morris

Based on Morris (2007), d_{ppc2} is an effect size estimator specifically designed for pretest-posttest-control (PPC) group designs. Morris specifies that this estimator offers high precision and robustness to assumption violations.

The formula goes as follows:

$$d_{ppc2} = c_P \left[\frac{(M_{post,T} - M_{pre,T}) - (M_{post,C} - M_{pre,C})}{SD_{pre}} \right]$$

where: - $M_{post,T}$, $M_{pre,T}$: Posttest and pretest means for the **treatment group** - $M_{post,C}$, $M_{pre,C}$: Posttest and pretest means for the **control group** - SD_{pre}: Pooled pretest standard deviation across both groups:

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

- c_P: Small-sample bias correction factor:

$$c_P \approx 1 - \frac{3}{4(n_T + n_C - 2) - 1}$$

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., Accolla, E., et al. (n.d.). Aarts, AA et al.(2015). Estimating the reproducibility of psychological science. Science 349, aac4716–aac4716. DOI. *Modelling Structure and Function of the Human Subcortex*, *52*, 299.

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.

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

Authority, E. F. S. (2014a). EFSA guidance document for evaluating laboratory and field dissipation studies to obtain DegT50 values of active substances of plant protection products and transformation products of these active substances in soil. *Efsa Journal*, *12*(5), 3662.

Authority, E. F. S. (2014b). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, *12*(6), 3734.

Banks, G. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions about questionable research

practices in the field of management: A guest commentary. In *Journal of Management* (1; Vol. 42, pp. 5–20). Sage Publications Sage CA: Los Angeles, CA.

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

Burgman, M. A. (2015). *Trusting judgements: How to get the best out of experts*. Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication." *Perspectives on Psychological Science*, *17*(5), 1490–1505.

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

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203.

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis." *American Psychologist*, *77*(3), 453.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *17*(1), 69–78.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*,

3(4), 456–465.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.

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.

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

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

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

Hall III, R. P. (2023). Replication and reproducibility and the self-correction of science: What can JID innovations do? *JID Innovations*, *3*(3), 100188.

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.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, *9*(1), 169–180.

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

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. , *62*(3), 221–230.

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

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

Lydick, E., & Epstein, R. (1993). Interpretation of quality of life changes. *Quality of Life Research*, *2*, 221–226.

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567.

Magnusson, K. (2023). A causal inference perspective on therapist effects. *PsyArXiv*. https://doi.org/XXX

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

Malich, L., & Munafò, M. R. (2022). Introduction: Replication of crises-interdisciplinary reflections on the phenomenon of the replication crisis in psychology. *Review of General Psychology*, *26*(2), 127–130.

Murphy, J., Caldwell, A. R., Mesquida, C., Ladell, A. J., Encarnación-Martínez, A., Tual, A., Denys, A., Cameron, B., Van Hooren, B., Parr, B., et al. (2025). Estimating the replicability of sports and exercise science research. *Sports Medicine*, 1–21.

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

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582–592.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748.

Nosek, B. A., & Lakens, D. (2014). Registered reports. In *Social Psychology*. Hogrefe Publishing.

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

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

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., & Rakow, T. (2006). The psychology of judgement under uncertainty. *Uncertain Judgements: Eliciting Experts' Probabilities*, 33–59.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618.

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

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for

a cumulative psychological science? A reply to götz et al.(2022). *Perspectives on Psychological Science*, *18*(2), 508–512.

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

Riesthuis, P., Mesquida, C., & Cribbie, R. (2024). *Statistical (non) significance≠(un) successful replication: The importance of the smallest effect size of interest*. OSF.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295.

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100.

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

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

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168.

Appendix A

Title of the appendix

You can write appendixes in the same way as chapters. Just add the class `appendix` to the header.

Be sure to add it as an appendix in `_quarto.yml` as well.