



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

Vibration of Effects (VoE): Assessing the Robustness of the Association Between Smartphone Use and Mood

Vibration of Effects (VoE): Valutazione della robustezza dell'associazione
tra uso dello smartphone e umore

Supervisor:

PROF. GIANMARCO ALTOÈ

Candidate:

GAIA BERTIATO

MATRICOLA: 2152803

Co-supervisor:

DR. MATTEO MANENTE

Anno Accademico 2025/2026

Acknowledgements

I would like to express my sincere gratitude to Prof. Gianmarco Altoè for his guidance and for teaching me the importance of breaking ambitious goals down into attainable steps. It was essential for me to be able to count on his constant encouragement and vision, which gave shape and value to this project.

I would also like to extend my heartfelt thanks to Dr. Matteo Manente for his constant availability and for the time he dedicated to explaining complex concepts in an accessible way. His contribution was crucial both to the completion of this thesis and to the practical application of the VoE method to real data.

I would also like to thank Nina for sharing this journey with me and for being by my side during moments of uncertainty: your support was essential in helping me reach this point.

A special thanks goes to my parents and siblings, who have always supported me and believed in my abilities from the very beginning; without you, I would not be who I am today. I am also deeply grateful to my friends for their unwavering support and understanding throughout these challenging yet rewarding years.

Last but certainly not least, I would like to thank Mattia, who has been by my side through the ups and downs of this wonderful journey.

Table of contents

Summary	5
1 Replication Crisis in Psychology	7
1.1 Replication Crisis	7
1.1.1 Definitions of Reproducibility and Replicability	9
1.1.2 Theoretical and Validation Crises	11
1.2 Main Causes of the Replication Crisis	13
1.2.1 Questionable Research Practices (QRPs)	13
1.2.2 Researcher Degrees of Freedom and Analytic Flexibility	15
1.3 Effect Sizes and Their Interpretation	17
1.3.1 Statistical and Practical Significance	17
1.3.2 Misinterpretations and Consequences	19
1.4 Implications for Scientific Credibility	20
1.5 Possible Solutions and Methodological Innovations	21
1.6 Aims of the Current Study	23
2 Effect Size	24
2.1 The Role of Effect Size in Statistical Inference	24
2.2 Standardized Measures of Effect Size	27
2.3 Type M and Type S Errors: Implications for the Accuracy of Effect Size Estimates	30
2.4 Recent Perspectives on the Use and Interpretation of Effect Size in Psychology	32
2.5 Toward Vibration of Effects (VoE) Analysis	34

3	Vibration of Effects (VoE)	35
3.1	The Garden of Forking Paths and the Need for Analytical Stability	35
3.2	The Multiverse Analysis Framework as a Response to the Replication Crisis	36
3.2.1	Exploratory Approaches to the Multiverse	38
3.2.2	Inferential Approaches to the Multiverse	40
3.3	Theoretical and Statistical Foundations of the VoE	42
3.3.1	Robustness and the Concept of Analytical Invariance	43
3.3.2	Statistical Formalization and the Distribution of Effects	44
3.4	Data Visualization and Interpretation: The Vibration Plot	46
3.4.1	The Janus Effect: A Case Study on α -Tocopherol	48
3.4.2	Vibration of Magnitude: A Case Study on Environmental Cadmium	49
3.5	Methodological Advantages and Limitations	51
3.6	Toward Standardized Robustness Reporting	52
4	Application of the VoE to the Smartphone Use and Mood Association	54
4.1	Ecological Momentary Assessment (EMA)	55
4.2	Smartphone Use and Mood: A Literature of Inconsistency	56
4.3	Rationale for Applying VoE	57
4.4	Method	58
4.4.1	Study Design and Participants	58
4.4.2	Data Collection Procedure	58
4.5	Variables	59
4.5.1	Momentary Mood	59
4.5.2	Smartphone Use	60
4.5.3	Covariates	61
4.6	VoE Framework and Specification Space	63
4.7	Results	65
4.7.1	Influence of Analytic Decisions	65
4.7.2	Volcano Plot	66
4.7.3	Specification Curve	67
4.8	Discussion	69

4.9 Conclusions	70
5 Conclusions	71
References	76

Summary

This thesis evaluates the *Vibration of Effects* (VoE) framework as a methodological tool for assessing analytical robustness, applying it to the association between smartphone use and mood using *Ecological Momentary Assessment* (EMA) data. The goal is to make a principled methodological contribution by demonstrating how the VoE framework can be applied, visualized, and interpreted in a realistic psychological research setting. The smartphone use-mood relationship serves primarily as a case study to illustrate the practical utility of this approach, rather than as an empirical end in itself.

Chapter 1 introduces the replication crisis as the motivating context for this work. A substantial proportion of published psychological findings fail to replicate under independent investigation, a problem driven by *Questionable Research Practices* (QRPs), *researcher degrees of freedom* (RDFs), and publication bias. In response, the *Open Science* movement has proposed a range of methodological reforms, including preregistration, registered reports, and multiverse approaches, that collectively shift the focus toward transparency and analytical accountability. This chapter establishes why such transparency is not optional but foundational to credible scientific inference, and positions VoE as one concrete methodological response to this challenge.

Chapter 2 examines the role of effect sizes in statistical inference, distinguishing practical magnitude from statistical significance. Standardized measures such as Cohen’s d and Pearson’s r are discussed alongside *Type M* and *Type S errors*, the *critical effect size* (CES), and the *smallest effect size of interest* (SESOI). The chapter concludes by introducing VoE as a robustness-oriented framework that moves effect size interpretation beyond the constraints of a single, arbitrarily chosen model.

Chapter 3 presents the theoretical and statistical foundations of VoE within the broader multi-

verse analysis landscape, distinguishing exploratory approaches from inferential ones such as *Specification Curve Analysis* (SCA) and *Post-Selection Inference in Multiverse Analysis* (PIMA). The vibration plot and its summary indices are then introduced and illustrated through two case studies from Patel et al. (2015), contrasting the directional instability of the *Janus effect* with the quantitative instability of vibration of magnitude.

Chapter 4 applies the VoE framework to EMA data from an Italian university subsample ($N = 126$), using the smartphone use-mood association as an illustrative empirical application. Across a predefined specification space of 100 models varying predictor coding, covariate sets, and model weighting, the partial correlation between smartphone use and mood was consistently positive, small, and statistically nonsignificant, with no Janus effect detected. The covariate set proved the most influential analytical decision, while model weighting had a negligible impact, illustrating how VoE can transparently map the sources of analytical variability and reveal which modeling choices drive observed differences in effect size estimates.

Taken together, this thesis argues that robustness analysis is not an optional supplement but a foundational requirement for credible psychological inference. By making the specification space explicit and transparent, VoE provides a concrete and practicable tool for promoting reproducibility, replicability, and analytical accountability in psychological science.

Chapter 1

Replication Crisis in Psychology

This chapter examines the replication crisis that has affected psychological research in recent decades. It begins with an overview of the crisis and the distinction between *reproducibility* and *replicability*, followed by an analysis of its main causes, including *Questionable Research Practices* (QRPs) and *researcher degrees of freedom* (RDFs). In this context, the chapter also discusses some key methodological sources of instability, such as analytic flexibility and the interpretation of *effect sizes*. The chapter then addresses the interpretation and potential misuses of effect sizes and their implications for the credibility of scientific findings. Finally, recent methodological innovations, including open science practices, *multiverse analyses*, and the *Vibration of Effects* (VoE) approach, are introduced as tools for examining the robustness of empirical results across alternative analytic specifications, concluding with the aims of the present study.

1.1 Replication Crisis

Over the past decade, the field of psychology has undergone a period of systematic change known as the *replication crisis*, wherein a substantial volume of empirical literature has been rigorously re-evaluated, revealing widespread failures to reproduce and replicate previously established findings (National Academies of Sciences & Medicine, 2019; Open Science Collaboration, 2015). *Reproducibility* is defined as the process by which independent researchers obtain the same results using the original researcher's data and analytical procedures; in contrast, *replicability* involves achieving

consistent findings when the same scientific question is tested using entirely new data (National Academies of Sciences & Medicine, 2019).

Replication is a central mechanism through which science builds cumulative and reliable knowledge (Hendriks et al., 2020; National Academies of Sciences & Medicine, 2019). Its primary function is to empirically evaluate the *reliability* of initial experimental findings, that is, the degree to which results remain stable and repeatable across time, observers, samples, or measurement instruments (American Psychological Association, 2018; Machery, 2020). Reliability is essential because it enables researchers to formulate predictable and generalizable claims about future observations (Schmidt, 2009). A successful replication is typically characterized by the observation of an effect that is statistically significant and consistent in both direction and magnitude with the original finding, thereby providing cumulative evidence for the phenomenon under investigation (Valentine et al., 2011). When such consistency is observed across independent studies, confidence increases that the effect reflects a genuine property of the population rather than an artifact of a particular sample or laboratory context (Ioannidis, 2005; Nosek et al., 2015). In this way, replication strengthens confidence in the *robustness* of empirical findings. A finding is considered *robust* insofar as it remains stable and it is not unduly sensitive to minor variations in experimental conditions (Klein et al., 2014).

A primary driver of the current crisis is the proliferation of *false-positive results* within the published literature (Pashler & Wagenmakers, 2012; Scheel et al., 2021; Shrout & Rodgers, 2018). A *false-positive result*, or *Type I error*, occurs when a researcher incorrectly rejects a true null hypothesis, thereby claiming an effect exists when the observed result was actually due to random chance (Pashler & Wagenmakers, 2012). Widespread Type I errors undermine the cumulative nature of science because they populate the literature with “discoveries” that do not exist, leading subsequent researchers to waste resources building upon a foundation of false premises (Pashler & Wagenmakers, 2012; Scheel et al., 2021).

The reality of this problem was demonstrated by the Open Science Collaboration (2015) project, which replicated 100 studies and found that only 36.1% yielded significant results ($p < .05$) in the same direction as the original findings (Open Science Collaboration, 2015). This high failure rate is particularly informative because replication studies often employ larger sample sizes to increase statistical power, thereby reducing the probability of *Type II errors*. A *false negative*, or *Type II error*, occurs when a researcher fails to reject a null hypothesis that is actually false, erroneously

concluding that no effect exists when a real phenomenon is present in the population (Pashler & Wagenmakers, 2012). Whereas the probability of a Type I error is controlled by the significance level α , high statistical power ensures that a replication study is sufficiently sensitive to detect the original effect if it is genuine. Consequently, when a well-powered replication fails to obtain a significant result, this outcome provides stronger evidence that the original finding may have been a false positive rather than a missed discovery. Such systematic failures to replicate have cast doubt on established theories and necessitated a fundamental shift toward more transparent research practices (Nosek et al., 2018).

The replication crisis has increasingly been interpreted as part of a broader *credibility crisis* affecting psychology and other disciplines, particularly in the social and biomedical sciences (Ioannidis, 2005; Open Science Collaboration, 2015). The replication crisis has moved beyond academic debate and is now widely discussed in public media, reflecting growing societal concerns about the trustworthiness of scientific research. For instance, more than a third of experiments in experimental economics fail to replicate (Camerer et al., 2016), and many reported findings in cancer treatment have proven irreproducible (Begley & Ellis, 2012). The persistent failure to achieve successful replication of key findings has triggered both a *theory crisis* (Eronen & Bringmann, 2021) and a *validity crisis* (Schimmack, 2021) in psychology, which will be discussed later.

1.1.1 Definitions of Reproducibility and Replicability

It is a common misinterpretation to treat reproducibility and replicability as synonyms; in reality, they represent distinct levels of verification. According to Nosek et al. (2022), *reproducibility* is the degree to which it is possible to recover the same results of a study using the original statistical analyses, code, and data. Consequently, a study may fail to be reproducible if the authors do not make their original data, code, or statistical analyses publicly available, or if the attempt to reproduce the results fails due to errors within the original study or the reproduction process itself (Nosek et al., 2022). Rather than gathering fresh data, reproducibility highlights the need for transparency and honesty about the original data processing decisions (National Academies of Sciences & Medicine, 2019).

In contrast, *replicability* involves obtaining consistent results across independent studies that are aimed at answering the same scientific question, but where each study collects its own unique data (National Academies of Sciences & Medicine, 2019). While reproducibility verifies the accuracy of

a specific analysis, replicability establishes whether a phenomenon is a reliable effect or merely a result of sampling or measurement error (National Academies of Sciences & Medicine, 2019; Nosek et al., 2022).

According to Machery (2020), a replication is fundamentally an experiment that resamples the “experimental components”—such as participants (experimental units), stimuli (treatments), or measurements—that are treated as random factors. *Random factors* represent elements drawn from a broader population, such as individual participants from a population or specific measurement tools from a population of available instruments. Consequently, the experimental elements subject to resampling and manipulation can include statistical units, independent variables, dependent variables, and the experimental setting.

Machery (2020) observe that while statistical units and settings are almost universally treated as random factors, other elements are often misclassified. Specifically, dependent variables (*i.e.* the measurements used to assess constructs) are frequently treated as fixed factors, despite often being sampled from a wider population of possible measurement tools. Typically, the independent variable remains the only true fixed factor in an experimental design (Machery, 2020).

Therefore, an experiment qualifies as a replication only if it faithfully reproduces the original study, with the exception of those components identified as random factors. If a researcher modifies the level of a factor that was originally treated as fixed, the study is no longer considered a replication; instead, it constitutes an investigation into a different relationship between variables rather than a verification of the original effect (Machery, 2020).

Traditionally, researchers distinguish between *direct replications* and *conceptual replications*. *Direct replications* aim to reproduce the original study as closely as possible, varying only the sample of participants, to determine whether the findings hold under the same methodological conditions (Nosek & Errington, 2020).

Conceptual replications, in contrast, test the same theoretical hypothesis using different methods, measures, or manipulations, allowing researchers to examine whether the underlying effect generalizes beyond the specifics of the original experiment (Nosek & Errington, 2020). Both types of replication play a critical role in scientific progress: direct replications provide a check on the reliability of specific experimental results, while conceptual replications evaluate the generalizability of theoretical claims (Machery, 2020; Patil et al., 2016). However, Machery (2020) argue that this distinction should be dissolved because both forms are essentially methods of resampling different

experimental components to verify the reliability of the original finding.

The failure to replicate a study often suggests that the initial finding was not an accurate reflection of the true population effect (Ioannidis, 2005; Nosek et al., 2022). These discrepancies generally arise from two distinct sources: random error and systematic bias. *Random error*, often called “noise,” involves unpredictable fluctuations during the measurement process, such as sampling variability or imprecise experimental control (Shadish et al., 2002). Because these errors are nondirectional, they lack a consistent bias and are just as likely to diminish a result as they are to inflate it. When an original finding is driven by this type of random noise, it is statistically unlikely to reappear in a subsequent, independent sample (Nosek et al., 2022).

In contrast, unrepresentative findings can also stem from systematic bias, such as Questionable Research Practices (QRPs), which differ from random noise because they consistently and artificially inflate the significance of a result, making it appear stronger than it truly is (L. K. John et al., 2012; Simmons et al., 2011). This problem is further compounded by publication bias, where journals tend to favor statistically significant results over null findings, creating a body of literature that does not reflect the full range of data collected (Head et al., 2015).

1.1.2 Theoretical and Validation Crises

While the replication crisis has primarily exposed the lack of consistency in empirical findings, it is increasingly viewed as a symptom of a much deeper credibility crisis rooted in both theoretical and methodological weaknesses. At the core of this issue lies a profound theory crisis, which refers to the lack of formalization, logical coherence, and falsifiability in psychological theories (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019). Many theories in the social sciences are articulated verbally, which leads to significant ambiguity and a “vagueness” that prevents precise predictions (Scheel et al., 2021).

A major driver of the theory crisis is the confusion between discovery-oriented research, which is exploratory in nature, and theory-testing research, which requires rigorous deductive logic (Oberauer & Lewandowsky, 2019). When researchers apply statistical tools designed for theory-testing to exploratory data without a solid theoretical foundation, they inadvertently inflate the rate of false positives (Oberauer & Lewandowsky, 2019). This inflation occurs because every individual statistical test carries a 5% risk of a Type I error. In exploratory settings, researchers often test many different relationships simply to see which ones appear significant (Oberauer & Lewandowsky,

2019). Without a theoretical anchor to limit these choices, performing multiple unplanned comparisons makes it statistically likely that at least one “significant” result will emerge from random noise (Wagenmakers et al., 2012). These random coincidences are then frequently mistaken for real scientific discoveries (Oberauer & Lewandowsky, 2019; Wagenmakers et al., 2012).

This is compounded by the fact that many psychological phenomena lack sufficient empirical evidence to support well-defined constructs, allowing for an indefinite number of competing theories to exist without a clear way to distinguish which are true (Eronen & Bringmann, 2021). This theoretical imprecision creates an environment where almost any data point can be post-hoc reconciled with a theory, effectively making the theory impossible to falsify (Oberauer & Lewandowsky, 2019). As Wagenmakers et al. (2012) argue, the concern is not merely the occurrence of isolated errors, but the systemic practice of post-hoc rationalization (Wagenmakers et al., 2012).

Parallel to the theoretical crisis is the validation crisis, which stems from a historical neglect of measurement quality and the widespread use of Questionable Measurement Practices (QMPs; Flake & Fried (2020)). Measurement serves as the fundamental link between hidden psychological constructs and observable data, yet researchers frequently overlook the validity of their instruments (Schimmack, 2021). The validation crisis is multifaceted, involving threats to internal, external, statistical, and construct validity. Internal validity concerns the ability to sustain causal inferences, while external validity relates to the generalizability of results, often limited by the use of non-representative WEIRD samples (Vazire et al., 2022). WEIRD is an acronym used to describe the demographic profile of the vast majority of participants in psychological research. It stands for: western, educated, industrialized, rich, democratic.

Construct validity, perhaps the most critical element, ensures that a measurement actually captures the intended psychological phenomenon. Unfortunately, many psychologists have traditionally settled for reliability—the mere consistency of a measure—as a substitute for validity, despite the fact that a reliable measure can still be fundamentally wrong if it does not measure what it claims to measure (Schimmack, 2021). This issue is worsened by a widespread lack of measurement transparency, as researchers frequently fail to report evidence for validity or justify their choice of scales in their published work. When these details are hidden, it creates significant uncertainty about whether the final claims are actually true and prevents the cumulative growth of scientific knowledge (Flake & Fried, 2020).

The term “validation crisis” specifically highlights how decisions made during the research pro-

cess, such as failing to report measurement details or ignoring low validity scores, undermine the credibility of psychological science. For example, discrepancies in how researchers define and operationalize constructs like “internet addiction” or “depression” show that without a shared and valid conceptualization, empirical results remain uninformative (Flake & Fried, 2020). Schimmack (2021) argues that validity cannot be a simple binary choice but must be treated as a dynamic, quantifiable process through causal modeling (Schimmack, 2021).

Ultimately, the theory and validation crises are deeply intertwined; without solid theories, researchers cannot develop valid measures, and without valid measures, theories cannot be accurately tested or falsified (Eronen & Bringmann, 2021). To address this, current perspectives suggest moving toward the mathematical formalization of theories and ensuring greater transparency in measurement reporting (Flake & Fried, 2020; Oberauer & Lewandowsky, 2019).

1.2 Main Causes of the Replication Crisis

1.2.1 Questionable Research Practices (QRPs)

The persistent prevalence of positive results in psychological literature suggests that the crisis is driven by structural external factors rather than mere random error (Fanelli, 2010; Scheel et al., 2021). Central to this issue is the *publication bias* inherent in scientific journals, which disproportionately favor innovative research with statistically significant outcomes. This environment incentivizes the *file-drawer effect*, a phenomenon where researchers discard non-significant or “negative” results, leaving them unpublished and inaccessible to the scientific community (Bakker et al., 2012; Fanelli, 2010; Rosenthal, 1979). Such systemic pressures have led to the widespread adoption of *Questionable Research Practices* (QRPs), defined as design, analytic, or reporting practices employed to present biased evidence in favor of a specific assertion (Banks et al., 2016). Although the community has been aware of QRPs for decades, recent data indicates their prevalence remains alarmingly high, with some estimates suggesting that over 50% of researchers have engaged in at least one QRP frequently (Gopalakrishna et al., 2022).

At the core of these practices lies the conventional application of *Null Hypothesis Significance Testing* (NHST). This procedure calculates the probability (p) of obtaining an effect at least as extreme as the one observed, assuming the null hypothesis is true (Head et al., 2015; Lakens, 2021).

The over-reliance on the arbitrary threshold of $p < .05$ as a gatekeeper for publication has created a powerful incentive for researchers to exploit *researcher degrees of freedom* to cross this significance line. This concept refers to the inherent flexibility in the scientific process, specifically the many subjective decisions researchers make regarding which observations to include, which conditions to combine, and which statistical analysis to perform (Simmons et al., 2011). Because these decisions are often made after the data have been collected and inspected, this flexibility allows researchers to steer the analysis toward a statistically significant outcome, even in the absence of a true effect (Gelman & Loken, 2013; Simmons et al., 2011).

One of the most prominent QRPs arising from this incentive is *p-hacking*, or *selective reporting*, which involves manipulating data collection or statistical analysis until a non-significant result becomes significant (Head et al., 2015). This flexibility—which includes “optional stopping” (*i.e.* ending data collection once significance is reached), the arbitrary exclusion of outliers, or the *post-hoc* selection of specific variables—drastically inflates the *Type I error* rate. In some cases, these practices can raise the probability of a false positive to 60%, making it more likely to detect a non-existent effect than to find nothing at all (Simmons et al., 2011).

Another pervasive practice is *HARKing*, or *Hypothesizing After the Results are Known*, where researchers present a *post-hoc* hypothesis as if it had been formulated *a priori*. This form of academic deception masks the exploratory nature of a study, presenting a “good story” with a “happy ending” that satisfies editorial preferences for clean, purposeful narratives. HARKing betrays the fundamental principle of honesty and risks the development of unnecessarily complex, unfalsifiable theories built around false-positive results (Kerr, 1998). Closely related to these is *cherry-picking*, the selective reporting of only those findings or citations that support a researcher’s prior beliefs while overlooking contradictory evidence (Murphy & Aguinis, 2019). Driven by confirmation bias, cherry-picking produces distorted literature that misguides subsequent research and compromises the magnitude of meta-analytic results (Mayo-Wilson et al., 2017).

The extent of QRP prevalence has been a subject of intense debate. A landmark study by L. K. John et al. (2012) suggested that over 90% of psychologists had committed at least one QRP in their career, leading to the conclusion that these practices are the norm rather than the exception. However, Fiedler & Schwarz (2016) challenged these findings, noting that the original survey was ambiguous and failed to distinguish between intentional misconduct and legitimate, though poorly reported, research choices. Their replication of the study, using more precise questioning, found

that while QRPs exist, their actual frequency might be an order of magnitude lower than initially reported. Nevertheless, the discrepancy between traditional literature and transparent methods remains striking; for example, while 96% of standard papers report positive results, only 44% of Registered Reports—which are designed to eliminate publication bias and QRPs—reach significant conclusions (Scheel et al., 2021). Ultimately, the widespread engagement in these practices is not merely a matter of individual ethics but a byproduct of a competitive academic culture that prioritizes the quantity of publications over the quality and replicability of science (Head et al., 2015; Nosek et al., 2022).

1.2.2 Researcher Degrees of Freedom and Analytic Flexibility

A central mechanism underlying the replication crisis is the pervasive influence of *researcher degrees of freedom* (RDFs), defined as the multitude of discretionary decisions researchers make throughout the planning, execution, analysis, and reporting of a study (Simmons et al., 2011; Wicherts et al., 2016). These decisions arise at every stage of the research process and introduce analytic uncertainty, including choices related to sample size determination, data exclusion criteria, operationalization of constructs, selection of covariates, model specification, and statistical thresholds. Although individual analytic decisions may each be theoretically or methodologically defensible, the presence of multiple plausible alternatives introduces substantial analytic flexibility (Steege et al., 2016). Consequently, different specifications of the same dataset can produce different estimates of the *effect size*, that is a standardized estimate of the magnitude of a phenomenon in the population (Kelley & Preacher, 2012). Examining how these estimates vary across plausible analytic decisions is therefore important for assessing the robustness of empirical findings and the sensitivity of results to analytic choices (Del Giudice & Gangestad, 2021; Steege et al., 2016). In this sense, robustness also refers to the extent to which results remain stable across different plausible analytic specifications of the same data (Del Giudice & Gangestad, 2021; Steege et al., 2016).

This flexibility begins at the design stage. In psychological research, it has traditionally been common to determine sample sizes based on conventions within a subfield, practical constraints, or rules of thumb rather than formal *a priori* power analyses (Cohen, 1962). Likewise, key analytic decisions, such as the operationalization of variables, inclusion of covariates, or selection of statistical models, are often not fully specified in advance but are refined during the research process (Simmons et al., 2011; Wicherts et al., 2016). Although such practices are often shaped by feasibility

constraints and disciplinary norms rather than intentional bias, they introduce flexibility into the research process. In the absence of clearly specified *effect sizes*—defined as quantitative estimates of the magnitude of a phenomenon in the population, representing the strength of a relationship or difference and serving as the parameter researchers ultimately seek to estimate (Kelley & Preacher, 2012; Lakens, 2013)—and predefined sampling plans, studies become vulnerable to underpowered designs. In such contexts, statistically significant findings are more likely to reflect sampling variability and inflated estimates rather than stable population effects (Altoè et al., 2020; Loken & Gelman, 2017).

Analytic discretion expands further during data preparation, where decisions concerning outlier exclusion, transformation procedures, variable construction, and participant removal can substantially alter outcomes (Steen et al., 2016). The problem intensifies at the modeling and reporting stages, where researchers may select among multiple statistically defensible specifications but ultimately report only those yielding significant results (Del Giudice & Gangestad, 2021). Gelman & Loken (2013) conceptualized this cumulative decision-making process as the *garden of forking paths*, in which each analytic choice constitutes a branching point capable of leading to different empirical conclusions. Importantly, this multiplicity does not require deliberate *p*-hacking or intentional misconduct. Even well-intentioned researchers may unknowingly capitalize on chance when analytic decisions are influenced, consciously or unconsciously, by observed data patterns (Simmons et al., 2011).

RDFs are at the core of the replication crisis because such analytic flexibility increases the likelihood that statistically significant findings reflect selective analytic pathways rather than stable population effects, particularly in a publication environment that systematically favors novel and statistically significant findings over null results (Nosek et al., 2018; Pashler & Wagenmakers, 2012). This incentive structure reinforces selective reporting practices and amplifies the visibility of potentially inflated effects. When researchers make data-contingent decisions, they increase the probability of capitalizing on random variation, thereby inflating false-positive results (Simmons et al., 2011). A central statistical consequence of unchecked RDFs is therefore the inflation of the Type I error rate. Traditional *Null Hypothesis Significance Testing* (NHST) assumes a single, pre-specified analytic plan. However, when multiple alternative analyses are explored—such as testing different covariates, excluding participants, or fitting alternative statistical models—each analytic pathway represents an additional opportunity for a significant result to emerge by chance. Because

NHST does not account for this hidden multiplicity, reported p -values systematically underestimate the true likelihood of obtaining statistically significant findings. This occurs because the nominal p -value is calculated as if only a single confirmatory test was conducted, ignoring the additional opportunities for significance created by alternative analytic decisions (Simonsohn et al., 2020).

Beyond inflating statistical significance, RDFs can also exaggerate effect sizes. When researchers selectively report analytic decisions that yield statistically significant results, the resulting effect size estimates are likely to be upwardly biased, appearing larger and more stable than they truly are in the population (Simmons et al., 2011; Simonsohn et al., 2020). Such distortions can mislead theoretical interpretation and bias cumulative evidence (Steege et al., 2016), particularly in meta-analytic syntheses that rely on published estimates.

In the broader context of the replication crisis, these dynamics provide a direct mechanism linking individual analytic flexibility to large-scale replication failure (Gelman & Loken, 2013; Simmons et al., 2011; Wicherts et al., 2016). When original findings reflect one favorable pathway among many plausible alternatives, they are unlikely to generalize across independent samples or alternative analytic specifications (Del Giudice & Gangestad, 2021; Steegen et al., 2016). This implies that evaluating a single analytic specification may provide an incomplete picture of the underlying effect, making it important to systematically examine how results change across different defensible analytic decisions (Del Giudice & Gangestad, 2021). Replication studies, particularly those conducted with higher statistical power, often fail to reproduce such inflated results, not because the replication is flawed, but because the original effect was partially or wholly driven by analytic flexibility rather than a stable population parameter (Altoè et al., 2020; Loken & Gelman, 2017). Consequently, RDFs contribute not only to isolated false-positive findings but also to systematic replication failures, thereby undermining the reliability and cumulative progress of psychological science (Nosek et al., 2018; Simonsohn et al., 2020).

1.3 Effect Sizes and Their Interpretation

1.3.1 Statistical and Practical Significance

The replication crisis has revealed fundamental limitations in a psychological research culture that prioritizes statistically significant findings over accurate effect estimation, underscoring that sci-

entific progress depends not merely on detecting whether an effect exists, but on understanding its magnitude and stability (Kelley & Preacher, 2012; Lakens, 2013; Open Science Collaboration, 2015).

Within the framework of null hypothesis significance testing (NHST), statistical significance indicates the probability of observing data at least as extreme as those obtained, assuming the null hypothesis is true. However, a p -value does not convey information about the size or practical importance of an effect. To address this limitation, researchers rely on *effect size measures*, which provide a quantitative estimate of the magnitude of a relationship, association, or difference between variables or groups, independent of sample size. More formally, an *effect size* represents a standardized estimate of the magnitude of a phenomenon in the population, the parameter psychological research ultimately seeks to understand (Kelley & Preacher, 2012; Lakens, 2013). Whereas an effect refers to the underlying population phenomenon, an effect size quantifies its magnitude in a standardized metric, enabling interpretation and comparison across studies (Kelley & Preacher, 2012; Lakens, 2013).

Effect sizes typically fall into two broad families. The d family quantifies standardized differences between groups, whereas the r family quantifies associations between variables (Ellis, 2010). A widely used effect size in the d family is Cohen’s d , which expresses the difference between two group means relative to the pooled standard deviation (Cohen, 1988). In the r family, the correlation coefficient (r) represents the strength and direction of a linear association between variables.

The practical importance of an effect depends on theoretical and applied context rather than arbitrary numerical thresholds. Although Cohen’s conventional benchmarks (e.g., $d = .20$ = small, $.50$ = medium, $.80$ = large) provide general guidance, their interpretation requires domain-specific knowledge (Ellis, 2010). Effects labeled “small” may have meaningful real-world consequences when accumulated across individuals or applied to outcomes with substantial societal or clinical impact (Lakens, 2013). Conversely, larger statistical effects may have limited practical relevance if they do not meaningfully influence decisions or outcomes (Wilkinson & Task Force on Statistical Inference, 1999). Because these benchmarks were intended as heuristics rather than universal standards, effect sizes must be interpreted in light of substantive context and theoretical expectations (Cumming, 2013).

Emphasizing effect sizes shifts statistical inference from dichotomous decision-making toward the estimation of population parameters. Test statistics used to compute p -values can be transformed

into standardized effect size measures, and their uncertainty can be quantified using confidence intervals. A 95% confidence interval indicates that, under repeated sampling, 95% of similarly constructed intervals would be expected to contain the true population parameter (Greenland et al., 2016). Reporting effect sizes alongside confidence intervals therefore conveys both magnitude and precision, providing a more informative summary than statistical significance alone (Cumming, 2013; Kelley & Preacher, 2012; Lakens, 2013).

By prioritizing effect sizes and their precision, researchers can reduce the risk of overinterpreting statistically significant but practically negligible findings, one of the core issues highlighted by the replication crisis (Open Science Collaboration, 2015). This estimation-focused approach aligns with contemporary methodological reforms aimed at strengthening the robustness and cumulative progress of psychological science.

1.3.2 Misinterpretations and Consequences

Despite their conceptual advantages, effect sizes are often misinterpreted or misused. A common misconception is to equate statistical significance with practical importance. Small p -values may correspond to trivially small effects in large samples, where even negligible differences become statistically detectable. Conversely, non-significant results may still reflect effects of meaningful magnitude when statistical power is low (Lakens, 2013; Open Science Collaboration, 2015). Overreliance on p -values, without consideration of effect size and precision, perpetuates distorted inferences about the strength and relevance of empirical findings (Simmons et al., 2011).

Low statistical power further exacerbates misinterpretation. In underpowered studies, statistically significant results are more likely to be associated with inflated effect size estimates due to sampling variability. Gelman & Carlin (2014) described this phenomenon as a *Type M* (magnitude) error, in which the observed effect size overestimates the true population effect. In extreme cases, a statistically significant estimate may even point in the wrong direction: a *Type S* (sign) error. Both types of error are more probable under low-power conditions and contribute directly to replication failures (Altoè et al., 2020; Gelman & Carlin, 2014).

These distortions are often amplified by questionable research practices (QRPs). Techniques such as p -hacking, optional stopping, selective reporting, and analytic flexibility increase Type I error rates and inflate observed effect sizes (Simmons et al., 2011; Simonsohn et al., 2020). Because publication systems historically favor statistically significant findings, inflated or spurious

effects are more likely to be published, while null or smaller effects remain largely invisible (Fanelli, 2010; Rosenthal, 1979). This selective visibility generates unrealistic expectations regarding the magnitude and generalizability of psychological phenomena.

High-powered replication attempts frequently reveal smaller effect sizes or null results, exposing the instability of original estimates (Open Science Collaboration, 2015). The central concern is not whether an effect is statistically significant, but whether its estimated magnitude is robust and replicable. Exaggerated initial effects can lead to theoretical narratives and applied recommendations that cannot be sustained under more rigorous testing. Consequently, understanding effect sizes—their definition, estimation, and interpretation—has become central to methodological reform, as it is essential for assessing the robustness, practical relevance, and replicability of empirical findings (Kelley & Preacher, 2012; Lakens, 2013).

1.4 Implications for Scientific Credibility

The replication crisis has had far-reaching consequences for the credibility and public perception of psychological science. Large-scale replication initiatives that have revealed substantially smaller or non-replicable effects have prompted widespread concern about the reliability of published findings (Lakens, 2013; Open Science Collaboration, 2015). When widely cited findings fail to replicate, trust in psychological research may diminish among fellow scientists, practitioners, policymakers, and the broader public (Nosek et al., 2022; Open Science Collaboration, 2015). Because psychological research often informs public policy, education, organizational practice, and health interventions, uncertainty about the robustness of its evidence base may reduce the perceived legitimacy of the discipline and decrease reliance on its recommendations (Kelley & Preacher, 2012).

Beyond reputational costs, replication failures represent a substantial waste of scientific and societal resources. Funding agencies, research institutions, and individual investigators may invest time and money in pursuing, extending, or implementing findings that later prove inflated or spurious (Fanelli, 2010; Simmons et al., 2011). Low statistical power, inflated effect size estimates, and selective reporting distort the research record, creating inefficiencies in knowledge accumulation and diverting attention from more robust lines of inquiry (Gelman & Carlin, 2014; Simmons et al., 2011). In this sense, the replication crisis is not merely a methodological issue but also a structural and economic one that affects how resources are allocated within the scientific enterprise.

The implications are particularly consequential in applied domains such as clinical and health psychology. Therapeutic interventions, prevention strategies, and diagnostic tools are often justified on the basis of empirical evidence. When effect sizes are unstable or overstated, clinical guidelines and treatment decisions may rest on fragile foundations, potentially limiting patient benefit or misdirecting healthcare resources (Lakens, 2013; Rosenthal, 1979). For example, interventions shown to produce statistically significant but practically negligible effects may be adopted prematurely, leading to opportunity costs for patients and practitioners (Cumming, 2013). Although many evidence-based interventions remain well supported, the crisis has underscored the ethical responsibility to ensure that applied recommendations are grounded in adequately powered, transparently reported, and independently replicated research (Nosek et al., 2022; Open Science Collaboration, 2015).

Importantly, the replication crisis has also generated significant opportunities for reform. Methodological innovations such as preregistration, open data and materials, registered reports, and large-scale collaborative replications reflect a growing commitment to transparency and cumulative science (Lakens, 2013; Nosek et al., 2022). By prioritizing accurate effect size estimation, adequate statistical power, and reproducibility, psychological science can strengthen its epistemic foundations and restore public confidence. Rather than signaling the failure of the discipline, the replication crisis may ultimately represent a period of methodological self-correction—one that promotes more robust, transparent, and socially responsible research practices (Open Science Collaboration, 2015; Simmons et al., 2011).

1.5 Possible Solutions and Methodological Innovations

In response to the credibility crisis, psychological science has undergone substantial methodological reform aimed at strengthening transparency, reproducibility, and cumulative knowledge building. Central to these reforms is the Open Science reform, which encompasses principles and practices designed to increase accessibility, accountability, and rigor across all stages of the research process (Kathawalla et al., 2021). Open science initiatives include open-access publishing, the sharing of research materials, data, and analytic code, more transparent peer review procedures, and incentives for comprehensive reporting of results, including null findings (Munafò et al., 2017). By making research outputs more accessible and verifiable, these practices reduce barriers to replication and

enable independent scrutiny of empirical claims. Platforms such as the *Open Science Framework* (OSF) provide infrastructure for documenting projects, sharing materials, and archiving preregistrations, thereby facilitating collaboration and preserving a permanent, time-stamped research record (Foster & Deardorff, 2017). Although increased transparency may expose researchers to potential misuse of data, institutional safeguards and community norms can mitigate these risks while preserving the broader epistemic benefits of openness (Lewandowsky & Bishop, 2016).

Preregistration has emerged as one of the most visible and influential tools within the open science framework. By publicly specifying research questions, hypotheses, methodological procedures, and analysis plans prior to data collection, preregistration constrains undisclosed analytic flexibility and reduces the likelihood of outcome-dependent decision-making (Bakker et al., 2012; Kathawalla et al., 2021). Importantly, preregistration clarifies the distinction between confirmatory and exploratory analyses, thereby protecting exploratory work from being misrepresented as hypothesis-driven testing (Lakens, 2019). Although preregistration does not guarantee strong theoretical foundations or methodological quality (and poorly designed studies can still be preregistered), it enhances transparency regarding analytic intentions and improves the falsifiability of theoretical claims (Lakens, 2019; Nosek et al., 2018). Closely related is the *Registered Report* format, which shifts peer review to the design stage and grants in-principle acceptance before data collection begins (Chambers & Tzavella, 2022). By evaluating the theoretical rationale and methodological rigor independently of study outcomes, registered reports reduce publication bias and weaken the incentive to obtain statistically significant results at any cost. Emerging evidence suggests that registered reports are associated with improved methodological quality and reduced outcome reporting bias (Chambers & Tzavella, 2022).

Recent reforms not only promote transparency in research planning and reporting but also address the impact of researcher degrees of freedom on statistical conclusions. Analytical decisions, such as variable operationalization, exclusion criteria, covariate selection, or model specification, can meaningfully influence effect size estimates. The multiverse analysis approach systematically evaluates how different, defensible analytic pathways affect results by estimating effects across a comprehensive range of plausible specifications (Steenen et al., 2016). Rather than presenting a single analytic choice as definitive, researchers disclose the distribution of outcomes across alternative specifications, enabling readers to assess the robustness of conclusions. Closely related, the *Vibration of Effects* (VoE) framework systematically examines how effect size estimates fluctuate

across model specifications and analytic decisions, quantifying the sensitivity of findings to these choices (Patel et al., 2015). Approaches such as VoE make this variability explicit, often through innovative graphical representations that summarize the distribution of effect size estimates across many plausible specifications, providing a quick overview of how conclusions depend on analytic decisions (Del Giudice & Gangestad, 2021; Patel et al., 2015). When effect estimates vary substantially across reasonable specifications, conclusions should be interpreted cautiously. Multi-analyst collaborations extend this logic further by demonstrating how independent research teams analyzing the same dataset can reach different results, thereby empirically illustrating the influence of analytic variability (Aczel et al., 2021).

Together, these methodological innovations reflect a broader shift from selective reporting and dichotomous significance testing toward transparency, robustness evaluation, and the systematic accumulation of empirical evidence (Chambers & Tzavella, 2022; Lakens, 2019; Simmons et al., 2011). They acknowledge that scientific conclusions are shaped not only by data, but also by design choices, analytic decisions, and publication incentives (Altoè et al., 2020; Gelman & Carlin, 2014). While no single reform can fully eliminate bias, error, or theoretical limitations, the combined implementation of open science practices, preregistration, registered reports, replication initiatives, and robustness analyses, including multiverse and vibration-of-effects approaches, provides a coherent and evolving framework for enhancing the reliability, credibility, and long-term integrity of psychological science (Chambers & Tzavella, 2022; Foster & Deardorff, 2017; Kathawalla et al., 2021; Lakens, 2019).

1.6 Aims of the Current Study

This thesis examines the *Vibration of Effects* (VoE) framework by exploring its theoretical foundations and its methodological application in psychological research. In addition to evaluating the framework conceptually, this thesis includes a practical application of the VoE method to empirical data. The overall goal is to develop a clearer and more comprehensive understanding of the utility and limitations of VoE as a research tool in psychology.

Chapter 2

Effect Size

This chapter examines *effect sizes* alongside traditional *significance testing*, distinguishing the practical magnitude of an effect from the inferential information provided by p -values. The discussion first defines standardized measures within the r and d families, subsequently analyzing how design-related constraints, particularly underpowered studies, can lead to systematic *Type M* (magnitude) and *Type S* (sign) errors. By addressing contemporary perspectives on the use and interpretation of effect sizes, including the application of the *critical effect size* (CES) and the *smallest effect size of interest* (SESOI), the chapter highlights a shift toward prioritizing the cumulative impact and practical relevance of observed effects. This chapter concludes by introducing the *Vibration of Effects* (VoE) approach as a framework for examining how effect size estimates vary across alternative model specifications.

2.1 The Role of Effect Size in Statistical Inference

As established in the previous chapter, reliance on the p -value as the primary, and often only, criterion for statistical significance has contributed to challenges in the replicability of psychological findings (Open Science Collaboration, 2015). Although the p -value is used to assess statistical significance under the null hypothesis, it does not convey information about the magnitude or practical importance of the effect under investigation (Cohen, 1988). To address this, researchers utilize the *effect size* (ES). The ES quantifies the magnitude of a phenomenon on a standardized scale,

representing the strength of a relationship or the degree of difference between variables (Cohen, 1988; Kelley & Preacher, 2012).

To illustrate the practical relevance of effect sizes, Figure 2.1 presents a simplified and intentionally provocative example that highlights the discrepancy between statistical significance and practical importance. In this scenario, a study compares a group receiving a massage intervention to a control group. A researcher might claim that “massage improves sleep” because the results are statistically significant ($p < .05$); however, the data reveal that the massage group only slept 47 seconds longer than the control group. While the p -value confirms that this 47-second difference is unlikely to be a random accident, the effect size reveals that the difference is too small to be meaningful to a participant. This contrast between statistical significance and practical importance is central to modern methodology: a mathematical “success” should not be mistaken for a practically meaningful result (Kelley & Preacher, 2012; Lakens, 2013).

Illustration of the distinction between p-value and effect size

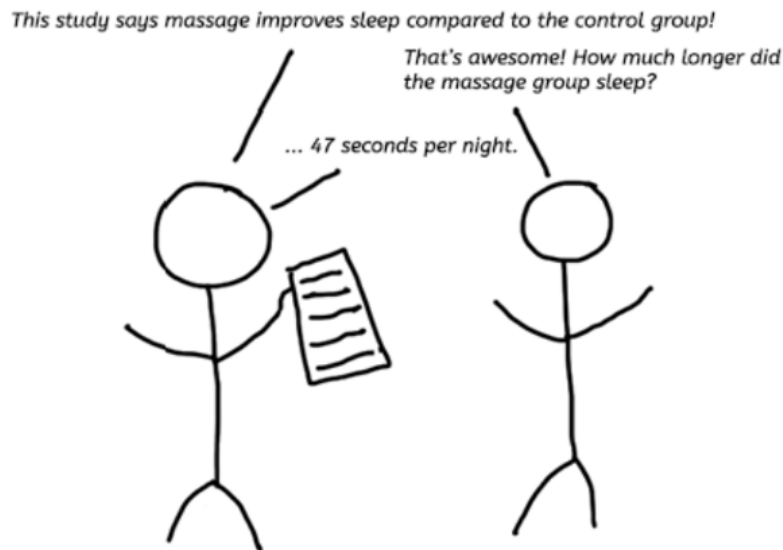


Figure 2.1: Note. In the example, a massage intervention produces a statistically significant difference in sleep duration ($p < .05$), yet the observed increase (47 seconds) is negligible in practical terms. The figure highlights how statistically significant results may correspond to trivial effect sizes (adapted from Altoè, 2023).

Although often reported together, the p -value and effect size serve fundamentally different purposes in statistical inference. The p -value quantifies the compatibility of observed data with the null hypothesis, representing the probability of obtaining results at least as extreme as those observed if the null were true. However, this metric is highly sensitive to sample size: in large samples, even trivial deviations from the null can reach statistical significance. In contrast, an effect size quantifies the magnitude of a phenomenon, that is, the strength of a relationship or difference between variables. It provides a standardized measure, such as a mean difference (d) or a correlation (r), that is not inherently a function of sample size, making it comparable across studies (Borenstein, 2009; Kelley & Preacher, 2012; Lakens, 2013).

Reporting both p -value and effect size conveys not only whether an effect exists, but also how large and meaningful it is within a specific theoretical or applied context (Cumming, 2013; Kelley & Preacher, 2012). This complementary approach ensures that statistical significance is interpreted alongside practical relevance. As emphasized by the American Statistical Association, no single statistic can replace scientific reasoning, and both p -values and effect sizes are necessary for a robust and informative evaluation of research findings (Wasserstein & Lazar, 2016).

Furthermore, focusing on effect size helps mitigate common misinterpretations that fuel the replication crisis. When studies are underpowered, statistically significant results often suffer from *Type M* (magnitude) errors, where the observed effect size is much larger than the true population effect, or *Type S* (sign) errors, where the direction of the effect is incorrectly identified (Altoè et al., 2020; Gelman & Carlin, 2014). Without a clear focus on the magnitude, these inflated estimates are published and create unrealistic expectations for future replications. By prioritizing effect sizes and their precision, often expressed through confidence intervals, researchers can shift from a binary “yes/no” decision-making process toward a more accurate estimation of population parameters. This transition is vital for distinguishing between findings that are statistically significant but trivial in magnitude - due to a large sample size - and those that reflect substantively meaningful psychological phenomena (Lakens, 2013; Open Science Collaboration, 2015).

Prioritizing effect sizes becomes particularly important in the context of the *Vibration of Effects* (VoE) framework. The replication crisis has revealed that different analytical choices, such as how to handle outliers or which covariates to include, can produce a range of different estimates from the same dataset, affecting both effect sizes and their statistical significance (Patel et al., 2015; Simmons et al., 2011). The *vibration* of effects refers to the variability in effect estimates that arises from

plausible alternative model specifications: when researchers systematically explore all reasonable analytical decisions, results can diverge substantially in magnitude and sometimes even direction (Patel et al., 2015). A focus on effect sizes allows assessment of how much the estimated magnitude of an effect changes across different models. If the effect size remains consistent across specifications, confidence in the robustness of the finding increases, whereas large fluctuations indicate results that are fragile and overly contingent on specific RDFs (Patel et al., 2015). A clear definition and reporting of effect size is therefore a foundational step for evaluating the robustness and practical relevance of psychological research.

2.2 Standardized Measures of Effect Size

Effect sizes can be expressed either in unstandardized units (e.g., raw mean differences) or in standardized form. Unstandardized effect sizes retain the original measurement units and are often more directly interpretable within a specific context. However, in psychological research, standardized effect sizes are frequently preferred because they allow comparison across studies, measures, and populations that may use different scales. For this reason, the present chapter focuses on standardized effect size measures.

Standardized effect size measures are generally categorized into two primary families based on the nature of the research design: the r family and the d family (Cohen, 1988). The r family focuses on the strength and direction of the association between variables. The most prominent metric, *Pearson's correlation coefficient* (r), represents the linear relationship between two continuous variables. Unlike the absolute covariance (σ_{XY}), which is expressed in original units and varies depending on the specific scale used, r is a standardized unit. It is defined as the ratio of the covariance to the product of the standard deviations of the two variables:

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (2.1)$$

In Equation 2.1, σ_{XY} represents the covariance between X and Y , while σ_X and σ_Y denote the respective standard deviations of each variable. By dividing the shared variance by the total variability, the metric is standardized and stripped of its original units of measurement. Consequently, a correlation of $r = .30$, for example, represents the same relative strength of association regardless of the underlying measurement scales (Cohen, 1988).

In contrast, the d family quantifies standardized differences between group means. Formally, the population effect size (denoted by δ) represents the theoretical “true” difference between population means divided by the common standard deviation (Cohen, 1988). In practice, researchers use the sample statistic d as an estimate of this parameter, calculated by dividing the observed mean difference by the pooled standard deviation (s_p):

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \quad (2.2)$$

In Equation 2.2, $\bar{X}_1$ and $\bar{X}_2$ represent the sample means, and s_p is the pooled standard deviation. This pooling process accounts for the variance and sample sizes of both groups to provide a more stable estimate of the population effect (Altoè et al., 2020). By standardizing the mean difference, Cohen’s d removes the original measurement units, allowing effect sizes to be expressed on a common scale and facilitating comparison across studies (Altoè et al., 2020; Cohen, 1988).

To assist researchers in interpreting these values, Cohen (1988) established conventional benchmarks where 0.2 represents a “small” effect, 0.5 a “medium” effect, and 0.8 a “large” effect. However, these values must be treated as heuristics rather than rigid scientific thresholds. Cohen (1988) cautioned against their universal application, noting that the magnitude of an effect is always relative to the specific research context, the difficulty of the task, and the theoretical expectations of the field. A “small” effect in a high-stakes clinical trial may carry more practical significance than a “large” effect in a controlled laboratory setting with limited real-world application (Altoè et al., 2020; Cohen, 1988).

While these benchmarks provide a useful categorical label, they remain somewhat abstract in terms of population impact. To provide a more intuitive understanding of what these differences represent, Cohen’s d can be transformed into probabilistic metrics such as the *Common Language* (CL) *effect size* and Cohen’s U_3 (Cohen, 1988; Ruscio, 2008). Figure 2.2 visualizes how these measures shift alongside increasing magnitudes of Cohen’s d .

Visualizing Group Overlap via Common Language (CL) and Cohen's U_3 Metrics

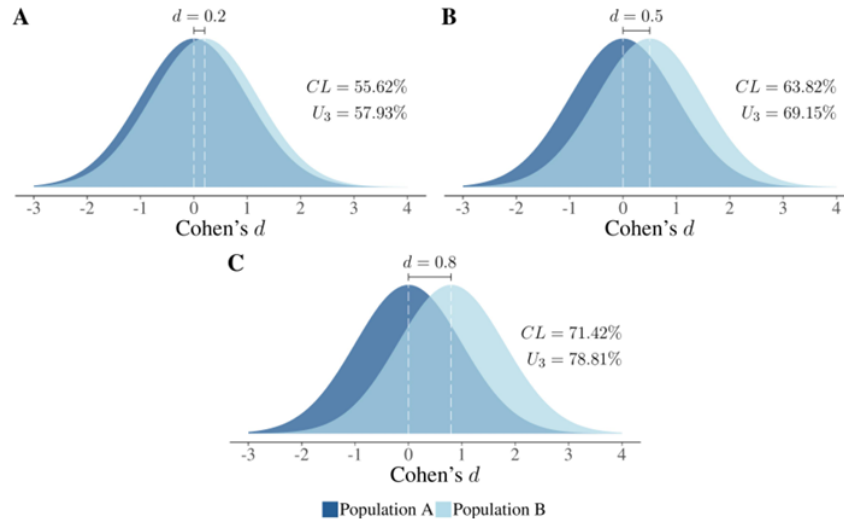


Figure 2.2: Note. This figure illustrates the practical separation between two populations (A and B) across three standard effect-size benchmarks. (A) Small Effect ($d = .2$): The distributions show substantial overlap; there is a 56% probability that a randomly selected individual from Population B will outscore one from Population A ($CL = .56$), with 58% of Population B exceeding the mean of Population A ($U_3 = .58$). (B) Medium Effect ($d = .5$): As separation increases, the probability of superiority rises to 64% ($CL = .64$), and 69% of Population B scores above the Population A mean ($U_3 = .69$). (C) Large Effect ($d = .8$): The distinction becomes more pronounced, resulting in a 71% probability of a higher score in Population B ($CL = .71$) and 79% of its members exceeding the mean of Population A ($U_3 = .79$) (adapted from Altoè et al., 2020).

The CL statistic represents the probability that an individual randomly selected from one group will score higher than an individual randomly selected from another. Similarly, U_3 indicates the percentage of one population that exceeds the mean of the other. These transformations translate abstract standardized differences into probabilistic or percentile terms, making effect sizes far more interpretable in applied and clinical contexts (Altoè et al., 2020; Cohen, 1988; Ruscio, 2008).

Finally, the reliability of standardized effect size estimates is heavily dependent on measurement accuracy. Although effect sizes are not defined as a function of sample size, their estimates must always be interpreted in light of sampling variability. In small samples, effect size estimates are inherently less precise and more unstable, increasing the likelihood of both overestimation and incorrect inference. This issue is compounded by measurement error. Because the standard deviation used in the denominator of Cohen's d reflects both true variability and measurement noise,

imprecise instruments can inflate the denominator and attenuate the observed effect size (Altoè et al., 2020). As a result, a seemingly small effect may reflect either a genuinely weak phenomenon or a stronger effect obscured by measurement imprecision. Importantly, low precision—due to small sample sizes or unreliable measurement—also increases the probability of Type M (magnitude) errors and Type S (sign) errors (Altoè et al., 2020; Gelman & Carlin, 2014). Consequently, effect size estimates should always be interpreted alongside considerations of sample size, measurement quality, and overall study precision.

2.3 Type M and Type S Errors: Implications for the Accuracy of Effect Size Estimates

While traditional null hypothesis significance testing focuses almost exclusively on the Type I error rate (α), a rigorous design analysis must also account for the risks of *Type M* (magnitude) and *Type S* (sign) errors (Altoè et al., 2020; Gelman & Carlin, 2014). A *Type M error*, or *magnitude error*, occurs when a statistically significant result provides a substantial overestimation of the true population parameter. In this scenario, the observed effect size is much larger than the actual effect, leading to a distorted view of the phenomenon’s practical importance. Conversely, a *Type S error*, or *sign error*, occurs when a statistically significant result carries the opposite algebraic sign of the true population parameter. This means a researcher may conclude that an effect is positive when it is actually negative, or vice versa (Gelman & Carlin, 2014). These errors are particularly prevalent when researchers attempt to detect subtle phenomena using noisy data or underpowered designs.

In these contexts, the criterion for statistical significance ($p < .05$) acts as a form of selective pressure, where only estimates characterized by extreme sampling variability in the direction of the effect are capable of surpassing the threshold required for publication (Gelman & Carlin, 2014). This pressure arises from the disparity between the true population parameter (δ) and the *critical effect size* (CES), which represents the minimum observed value necessary to reach statistical significance for a given sample size (Perugini et al., 2025). In underpowered studies, the critical effect size is typically much larger than δ . Since the true effect is insufficient to reach this threshold independently, a significant result can only occur when a large, positive sampling error is added to the true effect, thereby pushing the observed estimate past the critical value. Because the significance threshold filters out estimates that are representative of the true population parameter or those

that underestimate the effect, the published literature becomes biased toward these exaggerated values (Ioannidis, 2005; Perugini et al., 2025).

This systematic process defines the *exaggeration ratio*, which serves as a diagnostic metric to quantify the severity of the Type M error (Gelman & Carlin, 2014). The exaggeration ratio represents the factor by which the magnitude of an expected significant estimate exceeds the actual population parameter. When the ratio is near 1, the observed estimate is a stable representation of the population parameter. However, when the ratio is high (e.g., 4 or 5), the observed effect size is likely an artifact of chance rather than a reflection of the true phenomenon, regardless of a low p -value (Gelman & Carlin, 2014). Similarly, the risk of a sign error is quantified by the *Type S error probability*, which represents the likelihood that a statistically significant estimate carries the incorrect sign given the study’s design (Gelman & Carlin, 2014). In noisy or low power settings, the Type S error probability can be surprisingly high, suggesting that a non-negligible portion of significant findings may actually report an effect direction that is diametrically opposed to the truth (Altoè et al., 2020).

The identification of Type M and Type S errors highlights a fundamental tension in psychological research: the gap between achieving a statistically significant result and capturing a meaningful, stable estimate of human behavior. If the significance filter systematically prioritizes exaggerated magnitudes and unstable signs, then the mere presence of a low p -value is an insufficient indicator of scientific truth. This realization necessitates a shift in focus from the binary existence of an effect to a more nuanced evaluation of its size and reliability. Such a perspective challenges the traditional reliance on p -values as the sole gatekeeper of evidence and invites researchers to consider the practical weight of their findings within a broader theoretical and clinical framework. Recent methodological advances, which will be discussed in the following section emphasize that an effect size must be interpreted not just through its statistical probability, but through its practical utility and clinical relevance (Cumming, 2013; Funder & Ozer, 2019; Perugini et al., 2025). This transition toward prioritizing substantive magnitude over binary significance forms the basis for contemporary debates on the value of effect sizes in the psychological sciences.

2.4 Recent Perspectives on the Use and Interpretation of Effect Size in Psychology

Moving beyond the practice of simply confirming the presence of an effect has transformed how researchers evaluate empirical evidence (Cumming, 2013). As previously noted, the psychological research field is now prioritizing the stability and magnitude of an effect over its mere existence (Kelley & Preacher, 2012; Sullivan & Feinn, 2012). This methodological reform has generated new perspectives on both the technical use of effect sizes for validating research designs and their interpretation for evaluating practical importance.

Regarding the use of effect sizes, Perugini et al. (2025) emphasize the necessity of the *critical effect size* (CES) as a formal diagnostic tool for sensitivity analysis. Sensitivity analysis allows researchers to determine the minimum signal a study is capable of detecting given its specific constraints (Thabane et al., 2013). The CES represents the minimum observed magnitude required to reach statistical significance ($p < .05$) based on the sample size and alpha level (α). This value functions as a threshold for evidence; in high powered designs with large samples, the CES is small, allowing for the stable detection of subtle phenomena. Conversely, in underpowered studies, the CES is high, which creates a mathematical threshold that only inflated estimates can surpass. By using the CES as a benchmark, researchers can evaluate the credibility of an observed finding. A statistically significant result that only marginally exceeds a high CES is likely indicative of Type M error, suggesting that the estimate reflects random sampling noise rather than a stable population parameter (Perugini et al., 2025).

Furthermore, the CES provides a direct comparison to the *smallest effect size of interest* (SESOI), which represents the minimum magnitude a researcher considers theoretically or practically meaningful. For non significant results, the relationship between these two values determines whether a study was sensitive enough to provide evidence. If the CES is larger than the SESOI, the research design lacked the capacity to detect the very phenomenon it intended to measure. In such cases, a non significant result is technically uninterpretable because it fails to distinguish between a true absence of an effect and a mere lack of statistical sensitivity. If a study is essentially blind to meaningful effects, its failure to reach significance cannot be used to support the null hypothesis. Instead, the result represents a logical dead end where the data are insufficient to provide a reliable conclusion (Perugini et al., 2025).

Once the technical credibility of an estimate is established, the focus shifts to its interpretation. Funder & Ozer (2019) argue that the magnitude of a correlation is often undervalued due to a reliance on rigid, abstract benchmarks. For instance, an r of .30 is frequently dismissed as a “small” or “moderate” effect, a label that may fail to capture its actual practical utility. They suggest that instead of relying on such arbitrary thresholds, the strength of a relationship should be evaluated based on its concrete implications for the specific context under study (Funder & Ozer, 2019).

Specifically, the importance of such an effect becomes clearer when translated into a *binomial effect-size display* (BESD). The BESD is a conceptual tool that converts a correlation into a practical success rate by imagining a hypothetical comparison between two equal-sized groups (Rosenthal & Rubin, 1982). To maximize clarity, researchers typically illustrate this using two groups of 100 people as a standardized baseline. In this framework, the correlation r represents the absolute difference in percentage points between the success rates of two groups. Specifically, using the calculation $0.50 \pm (r/2)$, an r of .30 translates to a 30% difference: an individual in a high-scoring treatment group has a 65% chance of a positive outcome, compared to only a 35% chance for someone in a low-scoring control group (Funder & Ozer, 2019; Rosenthal & Rubin, 1982). As Funder & Ozer (2019) notes, this shift represents a substantial real-world impact; increasing a success rate from 35% to 65% nearly doubles the likelihood of a positive result (Funder & Ozer, 2019). This degree of change often represents the critical boundary between success and failure in clinical, educational, or professional interventions, demonstrating that even a “moderate” correlation can carry profound practical consequences.

This logic extends to group differences measured by Cohen’s d . Traditionally, a d of 0.20 is labeled “small”, yet its importance is clear when compared to medical benchmarks. In public health, the effect of a daily aspirin on heart-attack prevention has a d of approximately 0.06 (Sullivan & Feinn, 2012). While far below the traditional threshold for a small effect, it has meaningful practical implications since it prevents thousands of deaths annually when scaled across a population.

This perspective suggests that an effect’s value is not found in its absolute numerical size, but in its potential to accumulate over time or influence outcomes across large populations. A small effect that repeats daily, such as a minor personality trait or a subtle social influence, eventually creates massive differences in a person’s life trajectory in terms of health, social status, and overall well being. Similarly, an effect that appears tiny at the individual level can have profound societal

consequences when applied to millions of people. By integrating the CES to validate the use of an estimate with Funder and Ozer’s benchmarks for its interpretation, researchers can move beyond binary significance toward a more robust, consequence-based science (Funder & Ozer, 2019; Perugini et al., 2025).

2.5 Toward Vibration of Effects (VoE) Analysis

Beyond magnitude and utility, the credibility of an effect size depends on its robustness across varying analytical specifications (Patel et al., 2015). Vibration of Effects (VoE) describes the distribution of an estimate when a researcher systematically varies researcher degrees of freedom, such as the selection of covariates or data-cleaning criteria (Simmons et al., 2011; Wicherts et al., 2016). The primary utility of this approach lies in its capacity to reveal whether an effect is a stable phenomenon or merely an artifact of a specific analytical pathway (Simmons et al., 2011). A *robust* effect remains consistent across the majority of possible model specifications, whereas a *vibrating* effect, which fluctuates in direction or loses significance based on minor adjustments, suggests a lack of empirical reliability (Patel et al., 2015; Perugini et al., 2025). To evaluate this stability, researchers often employ visual representations, such as the *Janus plot*, to map the entire range of potential effect sizes and their corresponding *p*-values (Patel et al., 2015). These visualizations provide a transparent diagnostic for distinguishing a consistent discovery from a random result, a topic that, along with the theoretical and statistical foundations of VoE, will be examined in the following chapter.

Chapter 3

Vibration of Effects (VoE)

This chapter examines the *Vibration of Effects* (VoE) framework as a methodological response to the replication crisis and the analytical variability introduced by *researcher degrees of freedom* (RDFs), also known as the *garden of forking paths* (Gelman & Loken, 2013). The discussion begins by exploring the systematic instability of psychological findings across analytical choices and the transparency failures that conceal it. It then situates *multiverse analysis* as a key methodological response, distinguishing between *exploratory* and *inferential* approaches to the examination of analytical variability. Subsequently, the chapter details the theoretical foundations of VoE, specifically the concept of analytical invariance and the statistical formalization of effect distributions across the specification space. A guide to data visualization and interpretation follows, utilizing the *volcano plot* to analyze the density, sign, and significance of findings, with particular attention to the Janus effect. The chapter concludes by weighing the methodological advantages and limitations of the framework and proposing a movement toward standardized robustness reporting as a contemporary standard for scientific integrity.

3.1 The Garden of Forking Paths and the Need for Analytical Stability

In 2013, Gelman & Loken (2013) introduced the metaphor of the *garden of forking paths* to describe the multitude of potential analytical trajectories arising from researcher decision-making. Since

each analytical choice represents a distinct branching point in the analysis, a single study can lead to different, but equally plausible, conclusions based solely on the specific analytical path taken. Crucially, certain analytical choices are arbitrary yet equally justifiable, as they frequently occur in the absence of a singular, universally accepted theoretical framework or gold-standard measurement protocol. This scenario is frequent in psychology, where different ways of processing data and multiple model specifications can be considered plausible by scientific standards. The result is a multiverse of alternatively processed datasets and model specifications, where the reported finding is merely one implicit selection from an extensive array of potential outcomes (Steege et al., 2016).

Because each choice in the multiverse can lead to a different statistical result, any arbitrariness in the data construction and model specification phases is directly inherited by the final conclusion (Steege et al., 2016). A transparency problem arises when analytical decisions are made *post hoc*, without preregistration, allowing researchers to selectively report only those results that reach statistical significance — a practice known as *p-hacking* or *selective reporting* (Gelman & Loken, 2013). By presenting a single “optimal” result, the underlying analytical instability remains undisclosed, and the scientific community is left with a distorted and overly optimistic view of an effect’s true robustness (Simonsohn et al., 2020).

Aiming for analytical stability is therefore particularly vital for the field of psychology, where phenomena are complex and measurement models are rarely fixed. Because the constructs under study are often latent and can be operationalized in various ways, the credibility of an effect should be judged by its resilience across different specifications (Del Giudice & Gangestad, 2021). Establishing this stability ensures that findings are grounded in the global behavior of the data rather than in accidental analytical trajectories, which is essential for building a cumulative and reproducible science.

3.2 The Multiverse Analysis Framework as a Response to the Replication Crisis

Although the *multiverse analysis framework* was formalized after early contributions to the *Vibration of Effects* (VoE; Patel et al., 2015), it provides a crucial conceptual lens for addressing analytical variability in psychological research. In recent decades, the replication crisis has highlighted the fragility of psychological empirical findings, as many results fail to reach statistical significance

when subjected to independent reproduction and replication (Open Science Collaboration, 2015). While systemic factors such as publication bias are significant contributors (Fanelli, 2010), a more fundamental issue lies in the inherent instability of results across varying analytical choices (Steege et al., 2016). This instability is a direct consequence of RDFs, which refer to the multitude of discretionary yet defensible decisions made during data collection, coding, and statistical modeling (Simmons et al., 2011). In practice, RDFs allow for the construction of numerous alternative datasets and model specifications from the same raw data, such that different models can yield contradictory conclusions regarding the existence or direction of an effect, making it appear fragile (Steege et al., 2016).

In this context, the *multiverse analysis framework* has emerged as a comprehensive response to traditional reporting, shifting the focus from a single, arbitrarily chosen model to an examination of the distribution of outcomes across all theoretically plausible analytical paths (Steege et al., 2016). Specifically, multiverse analysis (MA) performs the study across the full set of datasets arising from all reasonable processing choices, while VoE evaluates how results change across different model specifications (Patel et al., 2015; Steege et al., 2016). This shift moves the analytical lens from a single, arbitrary path toward a systematic evaluation of all defensible analytical trajectories, acting as a modern extension of traditional sensitivity and outlier analysis (Steege et al., 2016). By acknowledging that a single analytical scenario cannot represent the complexity of research choices, this approach aligns with Open Science recommendations to provide a transparent, full accounting of how sensitive a result is to varying dataset configurations and model specifications.

To systematize the construction of the multiverse, Del Giudice & Gangestad (2021) propose a categorization of analytical decisions into three distinct types based on their theoretical and empirical justification.

Type E Decisions (Principled Equivalence) involve choices between truly interchangeable alternatives where there is no theoretical or logical reason to prefer one over another. For example, choosing whether to log-transform a variable or use a square-root transformation to address mild skewness often falls into this category. Because these choices are conceptually equivalent, a robust effect should be invariant across them.

Type N Decisions (Principled Nonequivalence) describe scenarios where one analytical choice is clearly superior to the others. This may occur when one measure has significantly higher reliability, or when one model specification correctly accounts for a known confounding variable while others

do not. Del Giudice & Gangestad (2021) argue that Type N choices should not be averaged or treated as equal in a multiverse; instead, the superior path should be prioritized.

Type U Decisions (Uncertainty) involve ambiguous scenarios where the correct path is unknown or depends on hidden assumptions. For instance, deciding on a specific cutoff for an extreme outlier when there is no consensus in the literature represents a Type U decision.

A central challenge identified by Del Giudice & Gangestad (2021) is determining when choices are truly arbitrary. If specifications are not equivalent (Type N misclassified as Type E), multiverse analysis can produce misleading results by obscuring meaningful effects within a mass of poorly justified alternatives.

Depending on the research goal, multiverse analysis can be categorized as either *exploratory* or *inferential* (Girardi et al., 2024). *Exploratory* approaches are primarily descriptive and aim to map the full range of possible outcomes in order to identify how and why results vary. In contrast, *inferential* approaches attempt to provide a formal statistical conclusion, specifically, to determine whether the overall pattern of evidence is stronger than what would be expected by chance. This distinction carries direct implications for how VoE findings should be reported and interpreted, a matter taken up in the following sections.

3.2.1 Exploratory Approaches to the Multiverse

Exploratory approaches to the multiverse serve as a descriptive map of the research process, systematically documenting how different analytical choices can lead to different outcomes (Steege et al., 2016). Rather than focusing on a single result, these methods chart the entire analytical landscape to reveal whether a finding is robust across many *forking paths* (Gelman & Loken, 2013). By treating variability as a source of insight rather than noise, these approaches provide a transparent overview of how findings react to choices in data cleaning or model selection. Within this framework, *Multiverse Analysis* (MA) and *Vibration of Effects* (VoE) provide the technical means to quantify this sensitivity, transforming hidden procedural choices into observable data.

Multiverse Analysis (MA) specifically addresses the problem of undisclosed preprocessing flexibility by focusing on the multiverse of datasets that can be generated from a single raw source (Steege et al., 2016). This method involves generating every unique permutation of analytical choices, such as different exclusion criteria or missing data treatments, that remains restricted to reasonable specifications supported by empirical or theoretical justification. As shown in Figure 3.1,

these outcomes are visualized through two complementary tools: a p -value frequency distribution (left), which illustrates the global stability of an effect, and a *specification grid* (right). The grid acts as a visual map linking each p -value to the specific set of operational decisions used to produce it. By making these analytical contingencies visible, MA identifies which data-coding decisions drive the final results and exposes the extent to which findings depend on RDFs (Steege et al., 2016).

Graphical representation of the results of a Multiverse Analysis

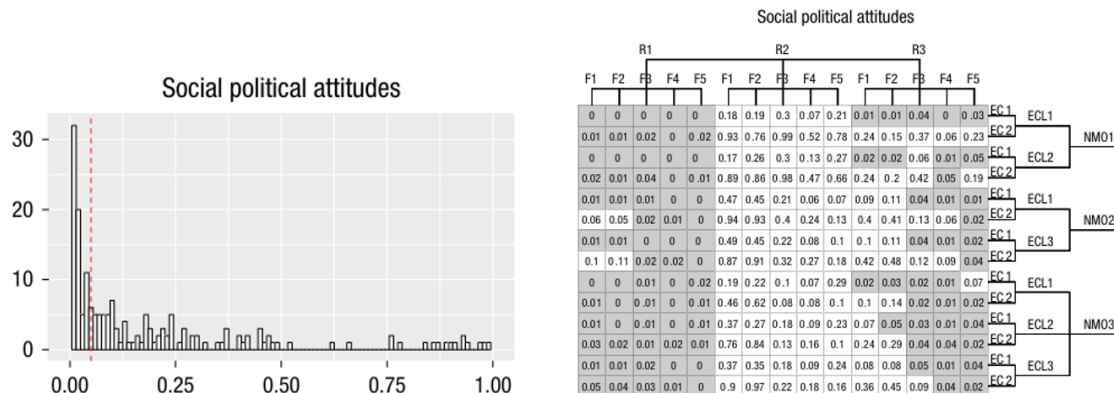


Figure 3.1: Note. On the left, the frequency distribution of the p -values resulting from the analyses conducted on each dataset of the multiverse; on the right, the grid of the possible combinations of data coding choices with respective p -values (combinations of choices with $p < .05$ are highlighted in gray). The acronyms (e.g., F1, NMO2, etc.) represent the diverse possible choices of data processing (adapted from Steege et al., 2016).

Complementing this focus on data preparation, *Vibration of Effects* (VoE) examines the fluctuations of results at the modeling stage (Patel et al., 2015). While MA explores how different versions of a dataset change a result, VoE investigates how different versions of a statistical model affect the outcome of a single dataset. In traditional research, authors frequently report only a single model specification, masking an informal process of trial and error, where multiple variations of an analysis are tested but only the most favorable result is published (Simonsohn et al., 2020).

This creates a problem of silent *multiple comparisons*, where the reported p -value is misleading because it does not reflect the true probability of a Type I error. When researcher test multiple models but only report the one with a significant result, the actual likelihood of finding a false positive is much higher than the nominal 5% threshold. In effect, the p -value is “artificially” low

because it fails to account for the numerous other forking paths that were explored but discarded (Gelman & Loken, 2013). Such *cherry-picking* of results, whether intentional or not, falls under the umbrella of *Questionable Research Practices* (QRPs) and significantly increases the risk of publishing false positives that lack genuine robustness (Ioannidis, 2005). The VoE framework systematically addresses this by estimating the distribution of the effect across predefined combinations of analytically defensible choices (Patel et al., 2015; Simonsohn et al., 2020).

A critical insight provided by the VoE is the identification of the *Janus effect*, named after the two-faced Roman god (Patel et al., 2015). This phenomenon occurs when the same dataset yields both positive and negative associations of similar magnitude, depending entirely on the chosen model specification, such as the specific set of control variables included. Such results reveal extreme fragility, suggesting that a finding might be an artifact of one specific analytical path rather than a reliable reflection of the phenomenon of interest. Ultimately, the wider the distribution of these effects, the more sensitive the finding is to the chosen statistical framework (Patel et al., 2015). By aligning with Open Science principles, these exploratory methods ensure that the stability (or instability) of an effect is fully disclosed, providing the necessary descriptive foundation for more formal statistical evaluations (Steenen et al., 2016).

3.2.2 Inferential Approaches to the Multiverse

Although exploratory mapping is essential for visualizing analytical instability, a descriptive account of effect variability is insufficient to distinguish genuine effects from the output of random noise (Girardi et al., 2024). Observing that effect estimates fluctuate across specifications leaves open the central inferential question: does the overall pattern of results reflect a true underlying phenomenon, or could it plausibly arise from random variation alone? Inferential approaches to the multiverse address this by introducing formal *hypothesis testing* across a predefined specification space, replacing qualitative impressions of instability with statistically grounded conclusions. For the purposes of the present thesis, these approaches are discussed insofar as they clarify the broader multiverse framework, while the focus of the present work remains the use of *Vibration of Effects* (VoE) to examine the robustness and variability of effect size estimates across model specifications.

Inferential methods share a common departure point, the evaluation of a *global null hypothesis*, which assumes that every analytical path in the multiverse captures random noise rather than a true effect. Rejecting this global null, that is, demonstrating that the distribution of results across

all plausible specifications is unlikely to have arisen by chance, provides stronger evidence than any single model result could offer in isolation. The two most prominent frameworks for this purpose are *Specification Curve Analysis* (SCA) and *Post-selection Inference in Multiverse Analysis* (PIMA).

Simonsohn et al. (2020) characterized *specification curve analysis* (SCA) as a hybrid methodology that pairs the descriptive transparency of multiverse visualization with a formal inferential framework. Once the full specification curve maps the ordered distribution of effect estimates across all reasonable analytical choices, the SCA framework uses resampling to generate a null multiverse. This null multiverse serves as a simulated baseline distribution under the assumption of a true effect size of zero. Researchers can then evaluate whether their empirical findings are statistically distinguishable from chance by comparing the observed curve to this baseline. Importantly, Simonsohn et al. (2020) emphasized that SCA assesses the collective robustness of the entire pattern of results across all specifications, rather than running separate inferential tests on each individual model.

However, SCA provides only weak control over the *Family-Wise Error Rate* (FWER), defined as the probability of committing at least one Type I error across the full family of tests conducted within the specification space (Girardi et al., 2024). When a large number of model specifications are evaluated simultaneously, this weak control allows the false positive rate to inflate, meaning that some specifications declared significant may not reflect true effects. For applications requiring inference about individual specifications rather than the curve as a whole, this limitation becomes consequential.

Post-selection Inference in Multiverse Analysis (PIMA) was developed precisely to address this limitation, offering a more statistically rigorous and flexible inferential framework (Girardi et al., 2024). PIMA provides strong FWER control, meaning that the probability of any false positive discovery across the predefined specification space is bounded at the nominal level ($\alpha < .05$) regardless of how many specifications are tested. This is achieved through selective inference procedures that account for the fact that researchers may search across specifications before making inferential claims, a process analogous to *multiple comparisons* correction, but tailored to the multiverse context. The result is a framework in which researchers can exhaustively explore the specification space to identify robust significant results while maintaining both inferential validity and good statistical power (Girardi et al., 2024).

Taken together, SCA and PIMA represent complementary rather than competing tools. SCA offers an accessible entry point for evaluating global robustness with strong visual interpretability,

while PIMA provides the formal framework needed when inferential precision at the specification level is required.

3.3 Theoretical and Statistical Foundations of the VoE

While *multiverse analysis* provides the broader framework for exploring analytical variability, *Vibration of Effects* (VoE) serves as a targeted diagnostic tool for quantifying how effect size estimates fluctuate across model specifications (Patel et al., 2015; Vinatier et al., 2025). Introduced by Patel and Ioannidis (2015), VoE emerged as a computational response to the reproducibility crisis in observational research. It represents a large-scale evolution of *sensitivity analysis*: whereas traditional sensitivity analysis relies on a researcher’s intuition to manually select a small number of alternative models, VoE automates this process to ensure an unbiased and exhaustive exploration of the specification space (Patel et al., 2015; Vinatier et al., 2025).

By estimating thousands of alternatives simultaneously, VoE identifies the extent to which arbitrary analytical choices dictate a study’s conclusions. VoE is particularly suited to observational research in fields such as psychology, where definitive causal models are rarely known in advance. Klau et al. (2023) distinguish between three primary sources of this instability: *model uncertainty*, concerning the inclusion or exclusion of specific covariates; *pre-processing uncertainty*, concerning the filtration and transformation of raw data; and *sampling uncertainty*, referring to fluctuations inherent to the specific observations included. Exploring these dimensions is critical because a finding that only emerges under a narrow set of conditions lacks the generalizability required for robust scientific claims (Patel et al., 2015).

From a statistical perspective, VoE is formalized as the joint distribution of an effect size, β , and its associated p -value across a large number of model specifications (Patel et al., 2015). Unlike a traditional analysis that yields a single point estimate, VoE produces a cloud of estimates representing the total potentiality of the data. The process typically begins with a base model capturing the primary relationship of interest. Researchers then systematically re-estimate this model using different combinations of covariates, variable transformations, and data filters to observe how the central estimate *vibrates* across the specification space (Vinatier et al., 2025).

By applying combinatorial analysis to generate every valid permutation of plausible analytical decisions, VoE ensures that the final distribution of results reflects the global behavior of the data

across all justifiable specifications, rather than the subjective preference of the researcher for one particular model (Patel et al., 2015; Steegen et al., 2016). This exhaustive process also enables the identification of the *Janus effect*, a phenomenon in which the same data set yields both positive and negative associations of similar magnitude depending entirely on the chosen model specification. The *Janus effect* will be discussed in detail in the following sections.

3.3.1 Robustness and the Concept of Analytical Invariance

By exhaustively mapping the specification space, VoE discloses the latent variability typically masked by a solitary analytical path, facilitating a transition from the traditional reliance on a single optimal result toward a transparent exploration of the distribution of potential findings (Simmons et al., 2011). At the core of this framework is the pursuit of *robustness*, defined as the consistency of a finding across the various arbitrary decisions made throughout the research process (Del Giudice & Gangestad, 2021).

In research practice, robustness is operationalized through the property of *analytical invariance*, which occurs when divergent but equally defensible model specifications yield substantially congruent results (Klau et al., 2023; Steegen et al., 2016). While robustness describes the conceptual stability of a finding, analytical invariance provides the formal empirical evidence for that stability, confirming that the signal is sufficiently resilient to survive analytical perturbation. If a psychological association is *robust*, it should remain reasonably stable across analytically defensible specifications that address the same substantive research question (Klau et al., 2023). A result that remains consistent across different specifications offers a level of evidence that an isolated point estimate cannot convey on its own, suggesting that the finding is unlikely to be contingent on the *forking paths* of a specific analytical workflow (Gelman & Loken, 2013; Patel et al., 2015).

It is important to note, however, that analytical invariance does not by itself guarantee *internal* or *external validity* (Shadish et al., 2002). Rather, a finding that persists across a multiverse of specifications strengthens confidence in its analytical robustness by demonstrating that the result is not overly dependent on a narrow set of analytical decisions, although it does not by itself establish causal validity or generalizability beyond the study design. This resilience suggests that the effect is not merely a byproduct of RDFs, indicating that the finding is a stable feature of the specific data under evaluation. This distinction is particularly vital in psychology, where constructs are often latent and operationalized through diverse methods. Without testing for analytical invariance,

a researcher may inadvertently publish a fragile result that cannot be replicated by independent teams who will inevitably make slightly different discretionary choices (Gelman & Loken, 2013).

Ultimately, the importance of analytical invariance lies in assessing the stability of findings. The goal is to determine whether an effect reflects a global property of the data or a local artifact contingent on a specific analytical trajectory. As Vinatier et al. (2025) note, this orientation aligns with the objectives of Open Science by making the fragility of empirical findings transparent. In this sense, VoE embraces the uncertainty inherent in the research process, using the distribution of potential outcomes as the benchmark for evaluating the reliability of a result (Patel et al., 2015).

3.3.2 Statistical Formalization and the Distribution of Effects

Moving beyond a purely conceptual understanding, VoE focuses on the generation of an effect size distribution in which the parameter of interest, typically a regression coefficient denoted as β , is estimated across a large array of model specifications (Patel et al., 2015). In this context, β represents the strength and direction of the association between a predictor and an outcome variable. Whereas a traditional study reports a single β as a static point estimate, the goal of VoE is to observe how that coefficient *vibrates* when the model is systematically altered across the predefined specification space.

In many large-scale epidemiological studies, β is expressed as a *Hazard Ratio* (HR), a metric particularly suited to longitudinal data in which the timing of an event is as informative as its occurrence. Because a single HR can be highly sensitive to the specific variables included in a model, the VoE framework introduces the *Relative Hazard Ratio* (RHR) to formally assess the stability of this risk estimate across specifications. Before detailing the formalization of the RHR, it is necessary to define the HR and the underlying concept of *hazard* on which it rests.

The HR is a measure used in survival analysis to compare the risk of an event occurring over time between two groups. It represents the ratio of the instantaneous event rates, known as *hazards*, in the two groups. An HR of 1.0 indicates no difference in risk between groups; an HR greater than 1.0 indicates higher risk in the exposed group; and an HR less than 1.0 indicates lower risk. The HR is typically estimated using the *Cox proportional hazards model*, which assumes that the ratio between the hazards of the two groups remains constant over time (Cox, 1972). For example, an HR of 0.70 means that, at any given point in time, the treated group has a 30% lower risk of experiencing the event compared to the control group. Importantly, the HR reflects a relative

rather than an absolute difference in risk.

The *hazard* itself can be understood as the rate at which events occur among individuals still at risk at a given moment. For example, if 100 individuals are initially at risk and 10 experience the event in the first month, the hazard is approximately $10/100 = 0.10$. If 90 individuals remain at risk in the second month and 9 experience the event, the hazard is again approximately $9/90 = 0.10$. This illustrates that the hazard reflects a conditional rate over time rather than a cumulative probability, making it sensitive to changes in the risk set at each observed time point.

The formalization process of VoE begins with the identification of a base model representing the primary association of interest. Researchers then define a set of potential adjusters, data filters, and variable transformations. Using combinatorial analysis, VoE executes every combination of these plausible choices to systematically map the predefined specification space (Vinatier et al., 2025). To quantify the resulting variability, Patel et al. (2015) introduced the RHR as a summary measure of the horizontal magnitude of the vibration. The RHR is calculated as the ratio of the 99th to the 1st percentile of the generated HR distribution:

$$RHR = \frac{HR_{99th}}{HR_{1st}} \quad (3.1)$$

In Equation 3.1, HR 99th represents the 99th percentile of the hazard ratio distribution, while HR 1st denotes the 1st percentile. Larger RHR values indicate greater variability across model specifications, whereas values closer to 1 suggest greater robustness and stability of the estimated effect.

An RHR close to 1.0 indicates a highly stable effect, meaning the risk estimate changes minimally across model specifications. Conversely, a large RHR signals that the magnitude of the estimated risk is heavily contingent on the specific analytical path taken. Complementing the RHR, the *Relative p-value* (RP) tracks the vertical fluctuation in statistical significance across the distribution of specifications. Together, these two metrics provide a transparent summary of analytical stability, establishing whether a finding constitutes a robust global property of the data or a fragile result contingent on specific modeling decisions (Patel et al., 2015).

3.4 Data Visualization and Interpretation: The Vibration Plot

The results of a VoE analysis are best understood through visualization rather than through numerical summaries alone. Because a VoE analysis can generate thousands, or even millions, of unique point estimates, traditional tables are insufficient to capture the resulting multimodal distributions or the specific influence of covariate shifting. To address this, researchers utilize the *vibration plot*, a specialized adaptation of the *volcano plot* commonly used in genomics (Patel et al., 2015).

The *vibration plot* is a graphical tool designed to represent the distribution of an effect size across the specification space, facilitating a comprehensive interpretation of analytical instability. By mapping the set of all plausible combinations of covariates, the plot addresses the limitations of single point estimation by revealing the latent distributional properties of the data. This visual approach allows for an immediate diagnostic of the data’s consistency: a tight, concentrated cluster indicates a robust signal, while a wide, scattered cloud warns of a result that is highly sensitive to the specification choices made during the analytical process.

Figure 3.2 illustrates two vibration plots obtained by Patel et al. (2015) comparing the association of mortality with Vitamin D (Panel A) and α -tocopherol (Panel D). While Vitamin D displays a robust, concentrated cluster situated entirely on one side of the null effect line, the α -tocopherol plot reveals a fragile *Janus effect* that crosses the null line. This crossover indicates that the result can flip direction or lose significance based solely on modeling choices, demonstrating that the finding is a byproduct of specific covariate selection rather than a stable property of the data.

Comparison of Robust (Vitamin D) and Fragile (α -Tocopherol) Associations

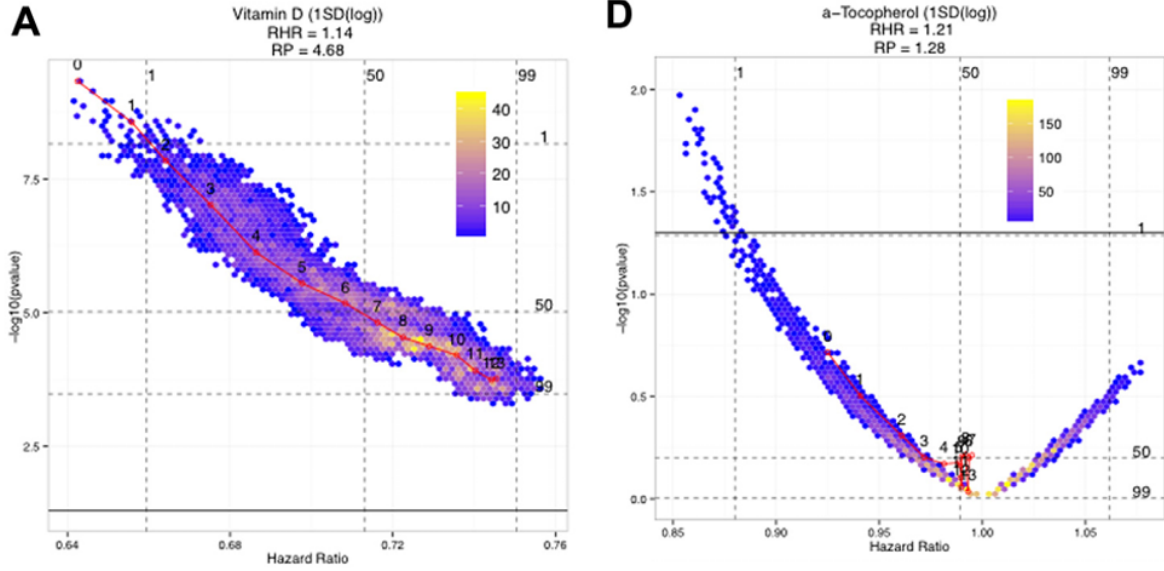


Figure 3.2: Note. Panel A (Vitamin D) displays a robust, concentrated cluster situated entirely on one side of the null effect line ($HR < 1.0$), indicating a consistent signal across the specification space. In contrast, Panel D (α -tocopherol) illustrates a Janus effect with a U-shaped distribution that crosses the null line ($HR = 1.0$), showing that the result can flip direction or lose significance based solely on the choice of covariates (adapted from Patel et al., 2015).

The distribution of results is evaluated through three formal dimensions: *density*, *sign*, and *statistical significance*. *Density* represents the frequency distribution of point estimates across the specification space; a high density cluster suggests that the result is an emergent, stable property of the dataset, whereas a sparse distribution indicates high specification sensitivity. *Sign* refers to the directionality of the coefficient, such as hazard ratios (HR) situated above or below the null value of 1.0, and *significance* tracks the proportion of the specification space that surpasses the alpha threshold ($\alpha = .05$).

To ground these formal concepts in a practical research context, the following section examines two concrete examples of VoE analysis taken from Patel et al. (2015). Patel et al. (2015) introduced VoE by performing a systematic sensitivity analysis of the *National Health and Nutrition Examination Survey* (NHANES) dataset. By subjecting various environmental and dietary variables to thousands of unique model configurations, the authors differentiated between *robust* associations, those that persist across the majority of the predefined specification space, and *fragile* associations

that manifest only under specific, non representative conditions.

3.4.1 The Janus Effect: A Case Study on α -Tocopherol

In their landmark study, Patel et al. (2015) examined the contested relationship between dietary intake of α tocopherol (Vitamin E) and all-cause mortality. While nutritional epidemiology often seeks to establish clear antioxidant benefits for such compounds, the authors’ VoE analysis revealed a state of extreme analytical instability known as the *Janus effect*. This phenomenon occurs when a single dataset yields diametrically opposed, yet statistically significant, conclusions based solely on model configuration.

As shown in Figure 3.3 (Panel D), the analysis highlights how a specific variable can inhabit different quadrants of the specification space. In this context, the vibration cloud is not contained within a quadrant, suggesting a single analytical outcome; rather, varying the set of adjusters pushes different valid models into the upper left quadrant (indicating a statistically significant protective effect) and the upper right quadrant (indicating a statistically significant hazardous effect). Because both extremes of this distribution exist above the $-\log_{10} p$ threshold, they both satisfy the criteria for significance ($\alpha = .05$), effectively allowing a researcher to *cherry-pick* a conclusion that supports a specific hypothesis.

The horizontal spread of the hazard ratios (HR) in Panel D reveals a distinctive bimodal or U-shaped distribution that crosses the null line ($HR = 1.0$). This pattern shows that the sign of the association is non-robust. By altering the set of adjusters, a researcher could derive models suggesting that α -tocopherol is either significantly protective ($HR < 1$) or significantly hazardous ($HR > 1$). Furthermore, the vertical axis, measured in $-\log_{10} p$, reveals that the maximum significance occurs at both poles of the effect size. This proves that the resulting conclusion is not a global property of the biological agent but a local artifact of the chosen statistical path. In this framework, specific combinations of covariates act as influential adjusters that pivot the association across the null. These covariates influence the resulting estimate, rendering the evidence directionally fragile and highly dependent on researcher discretion.

Graphical representation of a Janus Effect in a Vibration of Effects analysis

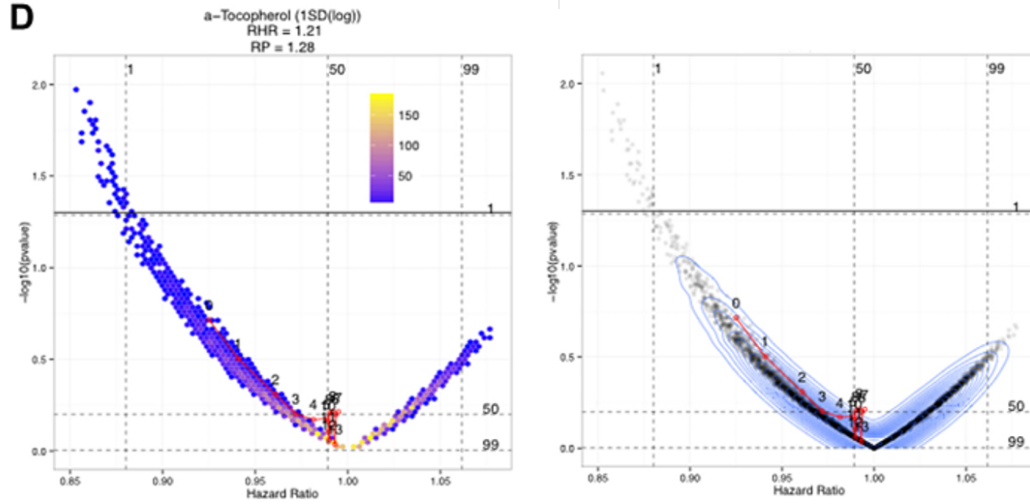


Figure 3.3: Note. This figure illustrates a Vibration Plot analyzing the association between dietary α -Tocopherol intake and mortality risk, providing an example of the Janus effect (adapted from Patel et al., 2015). Panel D displays a bimodal, U-shaped distribution that straddles the null line (HR = 1.0) on the X axis, indicating that the sign of the association is non-robust and can flip from protective to hazardous depending on model specification. The Y axis tracks statistical significance ($-\log_{10}p$), showing that both extremes of the vibration cloud ascend above the $p = .05$ contour lines. This visual split reveals that contradictory, yet statistically significant, conclusions can be derived from the same dataset through the selection of different covariate sets.

3.4.2 Vibration of Magnitude: A Case Study on Environmental Cadmium

Patel et al. (2015) also examined the association between environmental cadmium exposure and mortality risk. Cadmium is a toxic heavy metal and a Group 1 carcinogen. While cadmium is viewed as a significant health risk, its high correlation with smoking makes it a prime candidate for analytical instability. Because tobacco is a major source of both cadmium and mortality, it is difficult to determine if a perceived risk is driven by the metal itself or by confounding effects.

In contrast to the Janus effect, cadmium exposure reveals a finding that is directionally robust but quantitatively unstable. As shown in Figure 3.4 (Panel A), the cadmium vibration cloud stays entirely to the right of the null line (HR = 1.0), with results ranging from approximately 1.25 to 1.65. This indicates that the harmful association is a global property of the dataset as the sign never

flips. However, the horizontal spread quantifies a high *Relative Hazard Ratio* (RHR), suggesting that the intensity of the risk is highly sensitive to model choice.

The vertical position of the cloud tracks the *Relative p-value* (RP). While a large portion of the specifications remain statistically significant, the vibration toward the $p = .05$ threshold suggests that statistical confidence is highly contingent upon the model configuration. This plot identifies smoking status (Panel B) as the primary influential adjuster that shifts the cadmium vibration cloud. Because tobacco smoke is a major environmental source of both cadmium and mortality, it acts as a massive confounder.

When researchers include smoking as a covariate, it “absorbs” much of the mortality risk previously attributed to cadmium, systematically shifting the entire distribution of hazard ratios toward the null ($HR = 1.0$). As shown in Figure 3.4, this single analytical choice determines the study’s conclusion: without adjusting for smoking, cadmium appears as a severe, high magnitude risk ($HR = 1.6$), but once smoking is controlled for, the association remains significant but becomes much more modest ($HR = 1.3$). This shift illustrates that while the harmful effect of cadmium is a global property of the data (as it stays on the hazardous side of the null), its perceived intensity is highly sensitive to how researchers handle the “noise” of tobacco use.

Graphical representation of the results of a Vibration of Effects

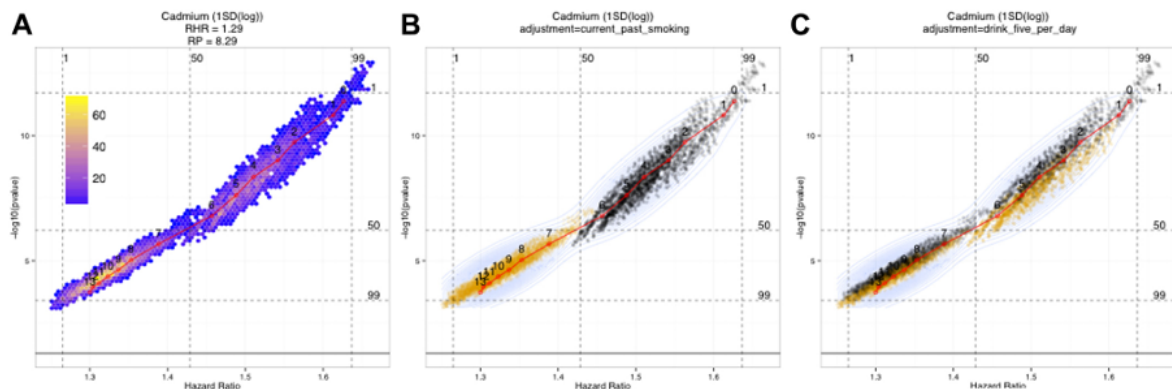


Figure 3.4: Note. This figure illustrates a Vibration Plot analyzing the association between environmental Cadmium exposure and mortality risk (adapted from Patel et al., 2015). Panel A provides a global view of the cloud, plotting the effect size (Hazard Ratio, β) on the X axis against its statistical significance ($-\log_{10}p$) on the Y axis. The horizontal spread quantifies the *Relative Hazard Ratio* (RHR), while the vertical spread tracks the *Relative p-value* (RP) relative to the $p = .05$ contour lines. Panels B and C highlight how the inclusion of specific adjusters, such as smoking status or alcohol consumption, can shift the distribution, revealing the association’s sensitivity to specific covariate sets.

The contrast between α -tocopherol and cadmium reveals that unstable results are not necessarily fragile in the same way. α -tocopherol exhibits a fundamental contradiction, in which the direction and significance of the association depend on researcher discretion, rendering the finding fragile. Conversely, cadmium represents a finding that is directionally robust but quantitatively unstable. The absence of a U-shaped distribution indicates a global association, even if the magnitude of the risk remains sensitive to influential adjusters such as smoking.

3.5 Methodological Advantages and Limitations

The transition from reporting a single statistical model to adopting the VoE framework represents a meaningful step forward for research transparency, though it involves both advantages and limitations. To determine whether VoE is appropriate for a specific study, researchers must balance its capacity for identifying analytical instability against the practical and interpretive challenges of high-dimensional modeling.

The primary advantage of VoE is the transparency it imposes on the research process. By mapping the distribution of results across the specification space, VoE reduces the opportunity for selective reporting by making the consequences of alternative analytical choices transparent and directly observable (Patel et al., 2015). Rather than providing a single point estimate, VoE tracks how a result fluctuates under different yet equally defensible conditions, offering a more comprehensive account of analytical uncertainty than is typically possible through fixed-specification modeling. Whereas a single p -value conveys evidence under one specific set of assumptions, a *vibration plot* illustrates whether statistical significance remains stable across a variety of alternative specifications. As Klau et al. (2023) suggest, this allows for a more transparent evaluation of whether a finding represents a robust association or is contingent on a narrow set of analytical choices. By shifting the focus from a single, optimal model to the global behavior of results across the specification space, VoE aligns with the broader objectives of Open Science and reproducible research (Steenen et al., 2016).

A significant limitation of this approach, however, concerns the risk of specification noise (Klau et al., 2023). When a researcher includes model configurations that lack sufficient theoretical or empirical justification, the resulting distribution of effect estimates can become excessively dispersed. In such cases, genuine signal becomes difficult to distinguish from analytical noise, and the vibra-

tion cloud may reflect researcher error in defining the specification space rather than true variability in the underlying effect. This risk underscores the importance of grounding every specification in substantive expert knowledge. A poorly constructed multiverse does not provide a more honest picture of reality; it provides a noisier one (Del Giudice & Gangestad, 2021).

VoE also presents practical challenges with respect to computational demand. Generating, running, and documenting thousands of unique model specifications requires considerable processing power and methodological expertise, which may constitute a meaningful barrier for researchers without access to adequate computational resources or technical training.

Finally, VoE can produce what might be termed an *interpretation paradox*: while presenting a wide distribution of results is more analytically honest, it can be difficult to translate into actionable conclusions for applied audiences such as clinicians or policymakers who often require a single, clear recommendation. When a finding is highly unstable — as in the Janus effect observed in the α -tocopherol study — it becomes particularly challenging to offer unambiguous practical guidance. As Del Giudice & Gangestad (2021) suggest, the challenge for the field is not only to generate the multiverse, but to develop principled methods for communicating complex distributional findings in ways that remain accessible and decision-relevant.

3.6 Toward Standardized Robustness Reporting

As the limitations of single-model reporting become increasingly evident, the implementation of VoE and multiverse approaches has brought attention to the need for a more standardized framework for robustness reporting. Scientific integrity, as Munafò et al. (2017) argue, requires more than raw data disclosure; it demands transparency about the stability of the claims derived from that data. In this spirit, future high-impact research may eventually incorporate a formal *robustness report* alongside every primary finding — a structured summary of how stable a result is across the specification space, analogous in function to the now-standard Data Availability Statement (Simmons et al., 2011; Steegen et al., 2016). Such a practice would establish the stability of a claim as a core metric of its scientific credibility, rather than a supplementary curiosity.

Rather than relying exclusively on the proportion of statistically significant specifications, robustness reporting should summarize multiple dimensions of stability. These include the direction of the effect across specifications, the range of effect size estimates, their practical relevance relative to

a *smallest effect size of interest* (SESOI), and the proportion of specifications yielding substantively similar conclusions. This multidimensional approach would avoid reducing robustness to another binary threshold and would preserve the estimation-focused logic that underlies VoE.

The practical adoption of these standards is increasingly feasible due to the development of dedicated software tools. *R* packages such as *multiverse* and *specr*, alongside recent algorithmic advancements by Vinatier et al. (2025), have substantially lowered the technical barrier to conducting large-scale sensitivity analyses. Careful statistical judgment, however, remains indispensable when defining the specification space and interpreting the resulting distributions.

Ultimately, the institutionalization of these methods may depend on a broader shift in researcher training. Encouraging the next generation of scientists to think in terms of analytical distributions rather than single paths, and to evaluate findings by their global stability rather than the significance of one chosen model, would represent a meaningful cultural shift in how psychological science is conducted and reported (Gelman & Loken, 2013). By embedding these practices in graduate education, the field could ensure that the inherent flexibility of the research process becomes a source of transparency rather than a mechanism for bias.

Chapter 4

Application of the VoE to the Smartphone Use and Mood Association

This chapter applies the *Vibration of Effects* (VoE) framework to empirical data on smartphone use and mood. It begins by introducing *Ecological Momentary Assessment* (EMA) as the data collection method, with attention to its methodological properties and suitability for studying real-time behavioral and affective processes. The literature on the smartphone use–mood association is then reviewed, documenting the persistent inconsistency of published estimates and motivating the need for a robustness-oriented analytic approach. The chapter proceeds with a description of the study design, participant sample, measures, and analytic strategy. Results are then presented, with particular attention to the distribution of effect sizes and the influence of analytic decisions on the estimated association. The chapter concludes with a discussion of the findings and an acknowledgment of the limitations of the analysis. The purpose of this chapter is not only to examine a specific empirical association, but also to demonstrate how the VoE framework can be applied, visualized, and interpreted in a realistic psychological research setting. The smartphone use–mood relationship serves primarily as a case study through which the methodological features of the approach can be illustrated.

4.1 Ecological Momentary Assessment (EMA)

Ecological Momentary Assessment (EMA), also known as the *Experience Sampling Method* (ESM), is an intensive longitudinal data collection technique. It is characterized by a high temporal density of repeated measurements captured over time to model within-person processes (Shiffman et al., 2008). This approach is considered *intensive* because it requires participants to complete assessments multiple times a day or week, resulting in a highly detailed dataset. This frequent data collection ensures temporal immediacy. By capturing self-reports in real time or near real time, EMA minimizes the retrospective memory biases that systematically distort conventional survey data (Csikszentmihalyi & Larson, 1987).

The EMA tradition was formally developed in the late 1970s by Csikszentmihalyi and Larson. Their early work used signal-contingent beeper protocols to study adolescent experience (Csikszentmihalyi et al., 1977). Since then, the method has evolved substantially. It has moved from paper diaries and pager-based signals to fully digital, smartphone-based platforms. These platforms enable both subjective self-reports and passive behavioral sensing (Yang et al., 2019).

This technological evolution has greatly increased the ecological validity and scalability of EMA research. *Scalability* refers to the method’s capacity to be extended to larger, geographically diverse samples without a proportional increase in logistical costs. This capacity makes it particularly well suited to the study of rapidly changing phenomena such as mood fluctuations and smartphone use (Trull & Ebner-Priemer, 2013).

From a methodological standpoint, EMA data are structured hierarchically. Multiple observations are nested within each individual, who provides repeated measurements over time. Since responses from the same person share variance, standard regression models are inappropriate. *Multilevel modeling* (MLM) is required to separate within-person fluctuations from between-person differences (B. Nezlek, 2011). The MLM framework accounts for the correlation in a person’s repeated answers by analyzing each participant’s unique baseline and their specific pattern of change over time. This approach allows researchers to detect micro-level processes, which would likely be invisible in cross-sectional designs (Shiffman et al., 2008). For example, it can detect mood changes in the hour following smartphone use.

4.2 Smartphone Use and Mood: A Literature of Inconsistency

The empirical literature on smartphone use and mood is characterized by persistent inconsistency, a pattern of fragility that mirrors the broader replication challenges discussed in Chapter 1. Cross-sectional and survey-based studies have frequently reported negative associations between screen time and well-being (Twenge & Campbell, 2018); however, these findings have been challenged on multiple methodological grounds.

Orben & Przybylski (2019) applied specification curve analysis (SCA) across three large-scale datasets ($N = 355,358$) and found that digital technology use accounted for an average of 0.4% of the variance in adolescent well-being, an effect size comparable in magnitude to minor, benign daily behaviors such as regularly eating potatoes. Critically, effect estimates were highly sensitive to analytic choices, ranging widely depending on how variables were operationalized and which covariates were included.

Longitudinal and ecological momentary assessment (EMA) studies have produced similarly heterogeneous results. Sewall & Parry (2024) tracked digital technology use and mental health symptoms over multiple weeks in a sample of young adults and found no significant within-person association over time. Only a between-person effect emerged: although heavier users tended to report slightly worse well-being overall, daily or weekly increases in phone use did not predict corresponding drops in mood. The authors interpreted this pattern as potentially reflecting pre-existing well-being differences rather than a causal effect of smartphone use.

In contrast, Minich & Moreno (2024) conducted an EMA study with 253 adolescents and found that smartphone use was positively associated with concurrent mood, with longer usage sessions linked to better mood states.

A systematic review of smartphone-based EMA studies further identified heterogeneity across designs, contexts, and outcome measures (Parry et al., 2021). Most studies have focused on group-level averages while neglecting within-person variability, a limitation that reduces ecological validity and constrains interpretability (Parry et al., 2021; Trull & Ebner-Priemer, 2013).

This pattern of inconsistency highlights the specific utility of vibration of effects (VoE) analysis. Because the relationship between smartphone use and mood may vary significantly based on sample demographics, measurement techniques, and model specifications, VoE provides a necessary

framework to systematically identify and quantify these sources of variability.

4.3 Rationale for Applying VoE

As reviewed above, published estimates of the association between smartphone use and mood vary widely in both magnitude and direction. Prior specification curve analysis has suggested that this variability is at least partly attributable to analytic choices rather than genuine population-level differences (Orben & Przybylski, 2019). The present dataset, currently being utilized in the initial phases of a research project in collaboration with Professors Alberto Acerbi and Ivano Bison from the University of Trento (UniTN), offers an opportunity to examine the instability of this association systematically. By applying the VoE framework to an EMA sample, this study aims to map how minor variations in data-analytic decisions alter final results.

The EMA design introduces a set of decisions that could reasonably be made differently by different researchers. These include selecting which covariates to incorporate, such as demographic, academic, or personality variables; determining how to operationalize smartphone use predictor variables derived from passive tracking; and deciding how to handle data cleaning and observation weighting. Each decision represents a *forking path* in an *analytic garden* where small changes in model specification can meaningfully influence the final estimate (Gelman & Loken, 2013). VoE analysis renders this process transparent by allowing researchers to systematically estimate the association across a predefined space of defensible model specifications and to visualize the resulting distribution of effect sizes.

If the smartphone use-mood association is genuinely robust, the distribution of estimates across specifications will be narrow and concentrated on one side of zero. If the association is fragile, the distribution will be wide and may exhibit a Janus-faced pattern, meaning that some specifications yield positive effects whereas others yield negative effects (Patel et al., 2015).

Although the direction and magnitude of the smartphone use-mood association remain contested, existing studies have rarely applied comprehensive robustness-oriented methods within EMA data. This analysis therefore applies the VoE framework to a real-world psychological dataset in order to illustrate how vibration analyses can be constructed, visualized, and interpreted in practice. Although the smartphone use-mood association serves as the empirical case study, the primary aim is methodological: to demonstrate how VoE can be used to evaluate the robustness and stability of

effect size estimates across alternative analytical specifications.

4.4 Method

4.4.1 Study Design and Participants

Data for this study were drawn from an international multicenter project conducted under the WeNet initiative, which extended across four global pilot sites located in Italy, Denmark, Mongolia, and the United Kingdom. The present analysis focuses exclusively on the Italian subsample, recruited from the student population of the University of Trento (UniTN). Data collection took place over four weeks between November and December 2020.

Recruitment followed a two-stage design. In the first stage, a standardized screening questionnaire was administered to approximately 10,000 students. Those who expressed willingness to participate were eligible for the second stage, from which approximately 1,200-1,300 volunteers were available. A stratified random sample was then drawn, stratified by degree level (bachelor's vs. master's/doctoral) to maintain representational balance, yielding approximately 320 enrolled participants, each compensated with €20 per completed two-week block.

After data cleaning and the exclusion of cases with insufficient observation density, the broader cleaned dataset comprised 132 participants. The final analytic sample for the Vibration of Effects (VoE) analysis consisted of 126 participants with complete data across all variables included in the specification space. The sample was predominantly female (54.0%), with a modal age range of 19-24 years (approximately 80% of participants). Most participants were enrolled in bachelor's programs (61.1%), with the remainder in master's or doctoral programs (38.9%). Disciplinary backgrounds were heterogeneous: social sciences and humanities ($n = 41$), STEM fields ($n = 37$), economics and law ($n = 30$), and natural sciences ($n = 18$).

4.4.2 Data Collection Procedure

The longitudinal phase of data collection relied on a customized mobile application serving a dual function: passive background logging of smartphone use data and active Ecological Momentary Assessment (EMA) prompting. During the first two-week block, participants were prompted every 30 minutes; in the optional second block, prompts were reduced to 60-minute intervals. At each

prompt, participants reported their current affective state, social context, and physical location. Simultaneously, the background application continuously recorded which apps were accessed and the exact duration of each active session.

Across the broad Italian longitudinal sample, this protocol yielded a total of 136,107 nested momentary observation windows. To meet the specific requirements of the current Vibration of Effects (VoE) framework, these intensive longitudinal observations were cleaned, filtered for missing data, and collapsed by calculating person-level averages across the tracking period. This process resulted in a final analytic sample of $N = 126$ participants who provided complete case data across all 13 operationalized variables within the specification space.

Consequently, the current analysis relied strictly on these cross-sectional, person-level aggregates rather than the full time-series structure of the Ecological Momentary Assessment (EMA) data. Because individual longitudinal trajectories were collapsed into global individual means, within-person dynamics (e.g., daily fluctuations or lagged effects) could not be examined. The results of this analysis therefore speak exclusively to between-person robustness, identifying stable (or unstable) individual differences in smartphone usage and well-being rather than momentary, intra-individual processes.

It should be noted that this analytical choice was guided by the primary objective of this work, which is designed as an illustrative methodological application of the VoE framework to digital tracking data rather than an exhaustive, formal multi-level modeling of longitudinal dynamics. This study aims to illustrate how VoE analysis can map the impact of analytical choices at the person level. Future research can expand this methodological framework into more complex, multi-level designs.

4.5 Variables

4.5.1 Momentary Mood

Momentary mood was originally assessed at each intensive longitudinal prompt using a single-item Ecological Momentary Assessment (EMA) question: “What is your mood?”. Participants responded using an emoticon-based ordinal scale ranging from 1 (sad) to 5 (happy). For the purpose of the current Vibration of Effects (VoE) framework, these individual longitudinal responses were

aggregated to the person level by calculating each participant’s mean mood score across all their valid tracking windows (`mean_mood`). Within the stable analytic subsample ($N = 126$), the aggregated person-level mood scores ranged from 2.87 to 5.00, with an overall sample mean of $M = 3.88$ ($SD = 0.49$). Consistent with typical ecological tracking data, these aggregated scores exhibited a negative skew, indicating a general baseline shift toward higher positive affect across the sample.

4.5.2 Smartphone Use

Smartphone use was captured via automated background tracking, which recorded objective usage durations across the monitoring period. Overall phone activity was indexed by the total cumulative duration of active phone use in seconds (`total_real_phone_sec`), which was subsequently normalized to reflect daily behavioral patterns by calculating the mean daily usage seconds (`mean_real_phone_sec_day`). For the purpose of the Vibration of Effects (VoE) analysis, this metric was scaled into mean daily application use in hours (`mean_real_phone_hours_day`) and its natural log transformation (`log_phone_hours_day`). The log transformation was applied to mitigate right-skewness and attenuate the influence of extreme heavy users.

As illustrated in Figure 4.1, the sample ($N = 126$) exhibited substantial individual variability in digital engagement, with mean daily application use ranging from less than 1 hr to over 8 hr per day. Reflecting typical behavioral patterns in digital tracking data, the distribution of raw daily application hours across participants displayed a pronounced right-skewed profile. This justified the systematic inclusion and testing of log-transformed predictors within the specification space.

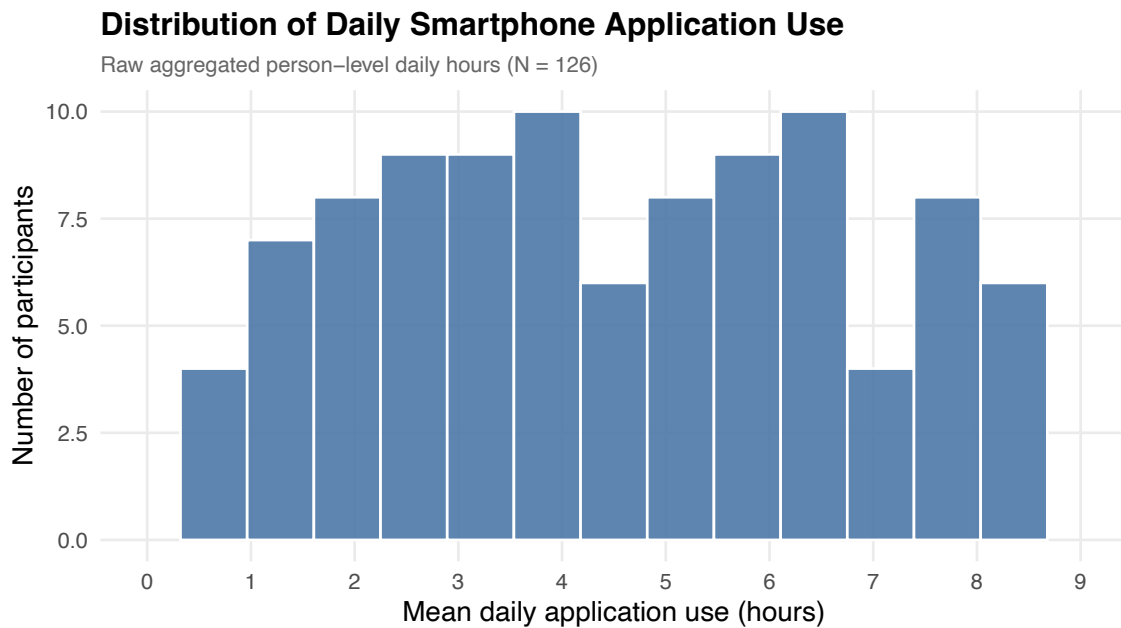


Figure 4.1: Distribution of raw mean daily smartphone application use across participants (hours/day). The distinct right-skewed profile highlights individual variability in digital engagement and motivates log transformation.

4.5.3 Covariates

Person-level covariates included sex, age cohort, academic degree level, and disciplinary field. Personality traits were assessed at baseline using the *Big Five Inventory* (O. P. John & Srivastava, 1999), yielding scores for Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. The covariates were organized into theoretically interpretable blocks. Rather than considering all possible combinations of these variables, a reasoned set of models was constructed, for a total of 25 different combinations. The chosen covariate sets are summarized in Table 4.1.

Table 4.1: Covariate Sets Included in the Analytic Specification Space

Set	Covariates included
Baseline	
1	No covariates

Set	Covariates included
Demographics	
2	Sex
3	Cohort
4	Sex + cohort
Academic background	
5	Degree
6	Department
7	Degree + department
8	Sex + cohort + degree + department
Personality (isolated)	
9	Neuroticism
10	Conscientiousness
11	Extraversion
12	Neuroticism + Conscientiousness
13	Extraversion + Agreeableness
14	Extraversion + Agreeableness + Conscientiousness + Neuroticism + Openness
Demo + Personality	
15	Sex + cohort + Neuroticism
16	Sex + cohort + Conscientiousness
17	Sex + cohort + Extraversion
18	Sex + cohort + Neuroticism + Conscientiousness
19	Sex + cohort + Extraversion + Agreeableness
20	Sex + cohort + Extraversion + Agreeableness + Conscientiousness + Neuroticism + Openness
Demo + Academic + Personality	

Set	Covariates included
21	Sex + cohort + degree + department + Neuroticism
22	Sex + cohort + degree + department + Conscientiousness
23	Sex + cohort + degree + department + Extraversion
24	Sex + cohort + degree + department + Neuroticism + Conscientiousness
25	Full model

4.6 VoE Framework and Specification Space

The Vibration of Effects (VoE) analysis was conducted on a complete-case dataset. Only participants with non-missing values on all variables in the specification space were included. This ensured that all model specifications were estimated on the same sample of 126 participants ($N = 126$), avoiding confounds between model specification and sample composition. Each specification was estimated as a linear model from which partial correlations were extracted as effect sizes.

The effect size of interest is the *partial correlation* (r_p) between smartphone use and mood. A partial correlation quantifies the association between smartphone use and mood after accounting for the covariates included in a given model specification. Positive values indicate that participants who use their smartphones more tend to report better average mood, whereas negative values indicate that participants who use their smartphones more tend to report worse average mood. Consequently, changes in the sign of the partial correlation across specifications would indicate a Janus effect, reflecting qualitatively different conclusions depending on the analytical choices made.

Overall variability across the specification space was summarized by two indices following Patel et al. (2015). The first index was the VoE spread (Δr_p), defined as the difference between the 99th and 1st percentiles of the partial correlation distribution. The second index was the *Relative Precision* (RP) *Index*, termed *Relative p-value* by Patel et al. (2015), which is the measure of vertical fluctuation in statistical significance across specifications, computed directly on the $-\log_{10}(p)$ scale.

The analyses were conducted using the *pima* package, an R package currently under development and associated with the forthcoming book *Multiverse Analysis in R: Exploratory and Inferential Approaches* (Altoè et al., in preparation).

The specification space was defined by three analytic decision nodes. The first decision node concerned predictor coding, comparing raw daily app use in hours or its log-transformed equivalent. The second decision node concerned the covariate set, which included 25 options organized into theoretically interpretable blocks, ranging from an unadjusted model to a fully saturated model including demographic, academic, and personality variables. The third decision node concerned model weighting, comparing an unweighted model or a model weighted by the number of valid mood observations per participant (`n_mood_valid`), giving greater influence to participants whose mood averages are based on more ecological momentary assessment (EMA) responses. As summarized in Table 4.2, the crossing of these three nodes yields a total of 100 specifications.

Table 4.2: Specification Space Defined by Three Analytic Decision Nodes

Decision Node	Alternatives	Number of Options
Predictor coding	Raw daily app use; Log-transformed daily app use	2
Covariate set	25 theoretically defensible covariate sets	25
Model weighting	Unweighted; Weighted by number of valid EMA mood observations	2
Total		$2 \times 25 \times 2 = 100$

The specification space considered in this application should not be interpreted as an exhaustive representation of all possible analytical choices, nor as a preregistered set of confirmatory analyses. Rather, it reflects a collection of theoretically defensible specifications designed to explore how the estimated association varies across plausible modeling decisions. Consistent with the logic of VoE, the goal is not to identify a single “correct” model, but to examine the stability of conclusions across a range of reasonable alternatives.

4.7 Results

Across all 100 specifications, the partial correlation between mean daily app use and mean mood ranged from $r_p = .034$ to $r_p = .097$, with a median of $r_p = .077$ ($M = .075$, $SD = .016$). The VoE spread was $\Delta r_p = .063$, and the relative precision index (termed relative p -value by Patel et al. (2015)) was $RP = 0.44$. All estimated effects were positive. No specification reached conventional statistical significance ($p < .05$) and no Janus effect was detected; the direction of the association did not reverse across any specification.

Figure 4.2 displays the distribution of effect sizes across all 100 specifications. The distribution is unimodal and right-skewed, with most estimates concentrated between $r_p = .06$ and $r_p = .09$. No estimate fell at or below zero.

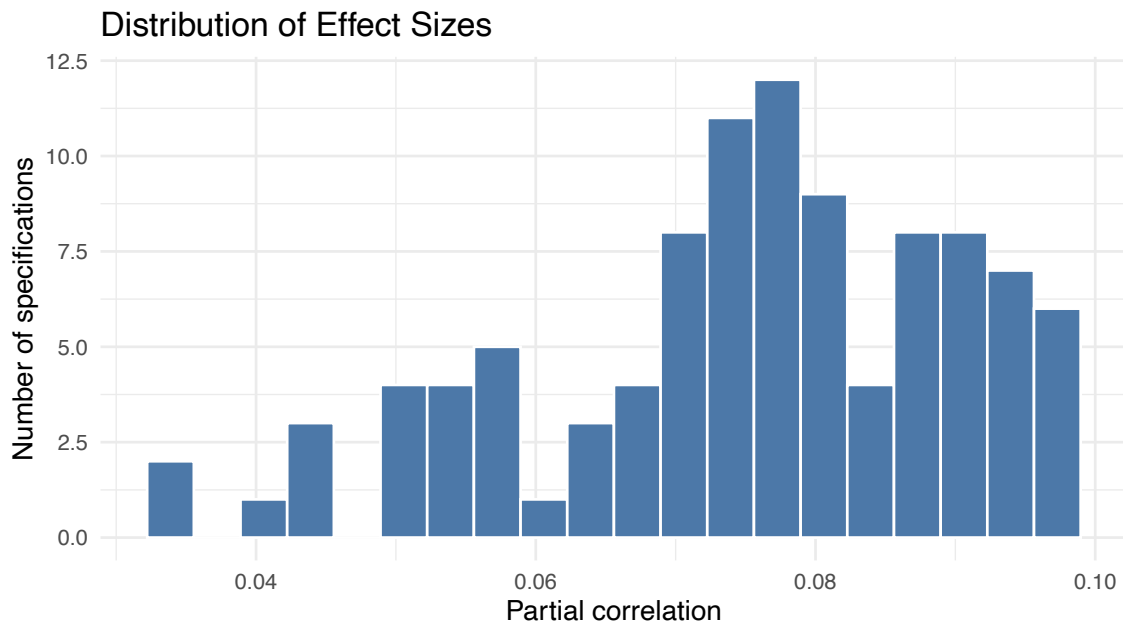


Figure 4.2: Distribution of partial correlations across 100 model specifications. Each bar represents the count of specifications falling within a given effect-size interval.

4.7.1 Influence of Analytic Decisions

The three analytic decision nodes differed in their influence on effect magnitude. Model weighting had the least impact. Median effect sizes were virtually identical for unweighted ($r_p = .077$)

and weighted models ($r_p = .077$). Participants who contributed more EMA responses did not systematically differ in the strength of the association.

Predictor coding was more consequential. Specifications using the log-transformed predictor produced larger effect sizes (median $r_p = .088$) than those using the raw predictor (median $r_p = .074$). This pattern is consistent with the right-skewed distribution of daily app use; by reducing the influence of heavy users, the log transformation prevents extreme outliers from distorting the trend and modestly strengthens the observed association.

The covariate set was the most influential decision node. Specifications including only neuroticism ($\bar{r}_p = .048$) or sex ($\bar{r}_p = .053$) yielded the smallest estimates, whereas fully adjusted models combining demographic variables with conscientiousness yielded the largest ($\bar{r}_p = .092$). Even this maximum observed effect remained *small* in absolute terms.

As shown in Figure 4.3, both median effect size and median statistical evidence increased modestly with the number of adjustment covariates—a pattern consistent with confound removal rather than inflated estimation.

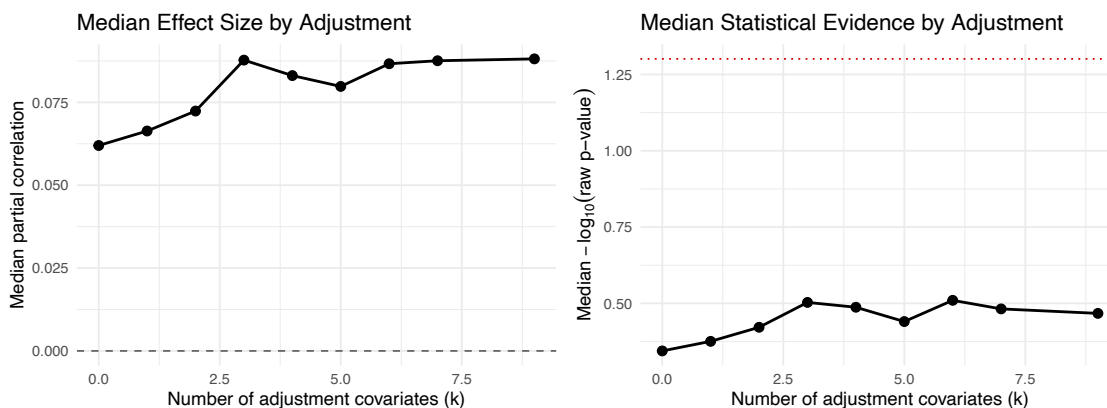


Figure 4.3: Median partial correlation (left panel) and median statistical evidence (right panel) as a function of the number of adjustment covariates (k). The dashed line in the left panel marks the null; the dotted red line in the right panel marks $p = .05$.

4.7.2 Volcano Plot

As shown in Figure 4.4, each model specification is plotted as an individual point, mapping the partial correlation on the x -axis against the statistical evidence ($-\log_{10}(p)$) on the y -axis. Within

the plot, a vertical dashed line marks the null hypothesis ($r_p = 0$), while a horizontal dotted red line denotes the conventional significance threshold ($p = .05$). The two separate panels contrast the raw predictor against its log-transformed equivalent, and the distinct point shapes differentiate between weighted and unweighted models. Across all 100 specifications, the data points clustered tightly within the positive region of the x -axis, with no analytical pathway approaching or crossing the threshold for nominal statistical significance.

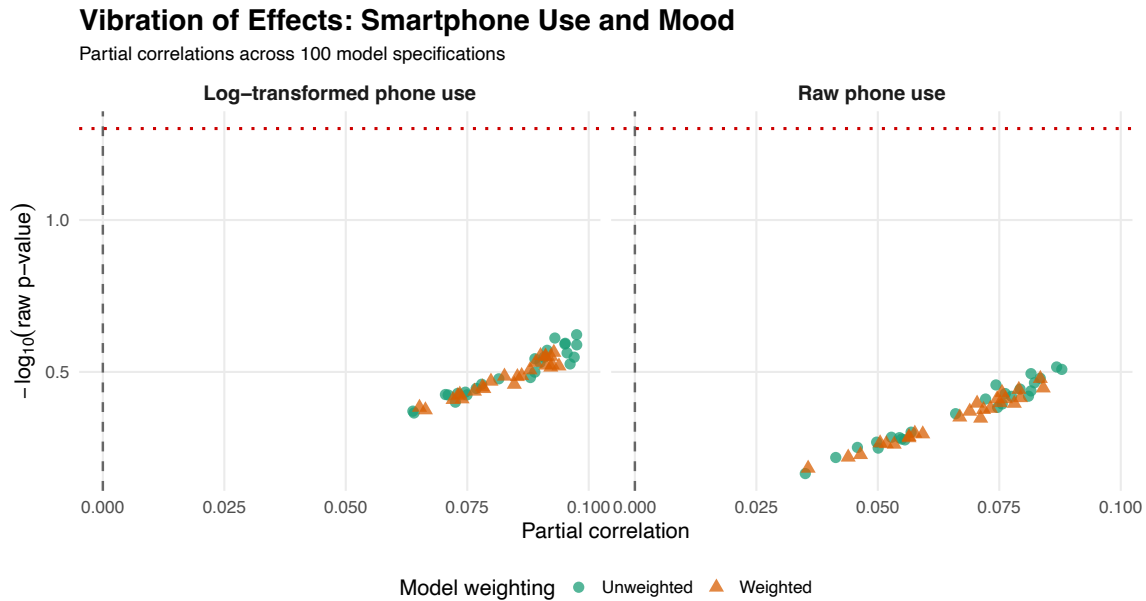


Figure 4.4: Volcano plot of partial correlations across 100 model specifications. Each point represents one specification. Color and shape distinguish model weighting; panels distinguish predictor coding. The vertical dashed line marks zero; the horizontal dotted red line marks $p = .05$.

4.7.3 Specification Curve

Figure 4.5 orders all 100 specifications by estimated effect size. The upper panel shows the partial correlation for each specification. The lower panel indicates which analytic choices characterize each specification. Blue tiles mark selected options; white tiles mark unselected options. Specifications with smaller effect sizes used raw predictors and personality-only covariate sets. Larger effect sizes used log-transformed predictors and demographic-heavy covariate sets. Weighting was distributed evenly, consistent with its negligible impact on estimates.

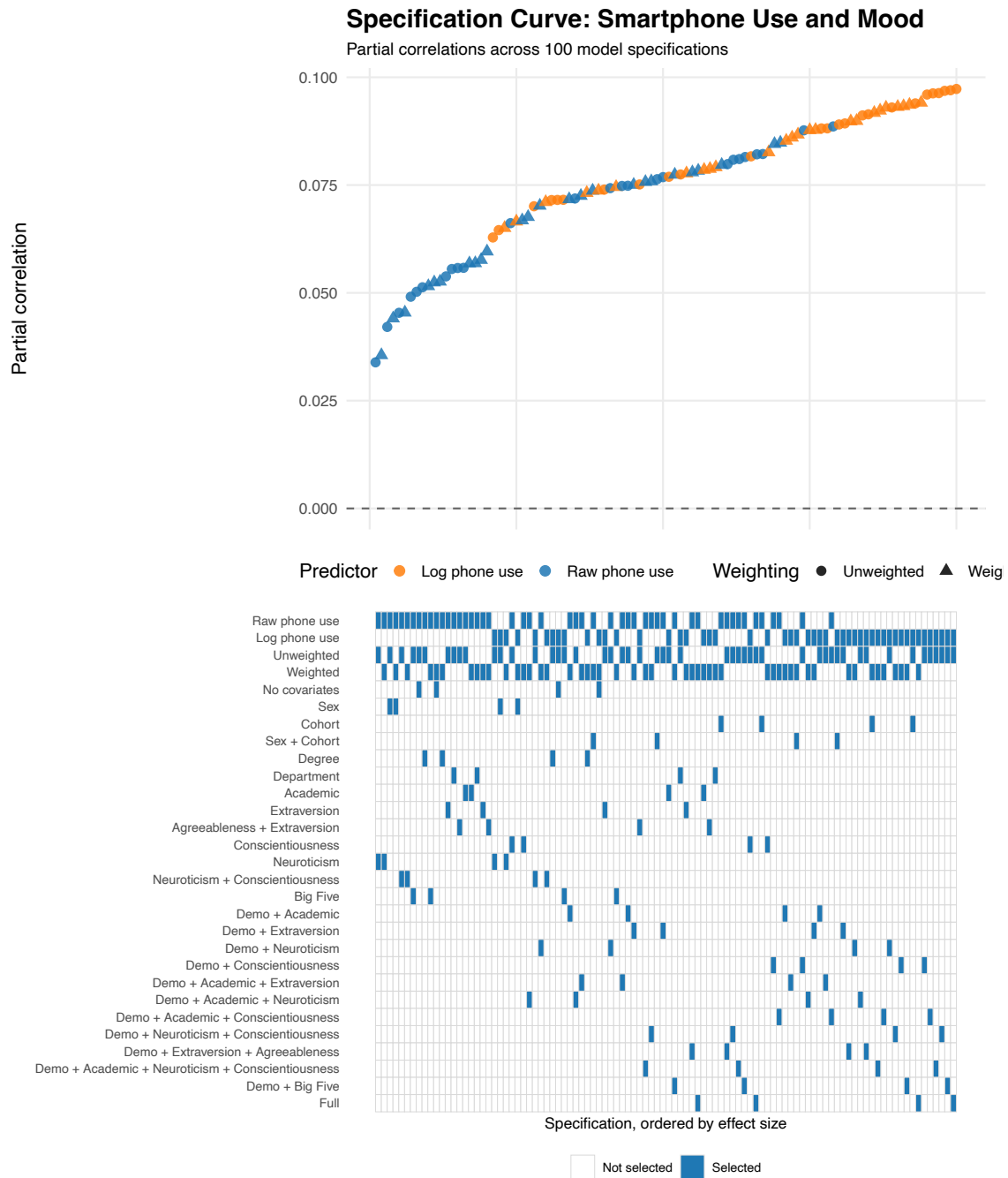


Figure 4.5: Specification curve for the smartphone use-mood association. The upper panel shows the partial correlation for each of the 100 specifications, ordered from smallest to largest. The lower panel indicates which analytic choices characterize each specification (blue = selected; white = not selected). Point color distinguishes predictor coding; point shape distinguishes model weighting.

4.8 Discussion

The VoE analysis yields a coherent and interpretable pattern. Across all 100 specifications, the association between mean daily smartphone use and mean mood was positive, small, and statistically non-significant. The partial correlation ranged from .034 to .097, with a median of .077, an effect size that is characteristically small for intensive longitudinal tracking but potentially meaningful over time (Funder & Ozer, 2019). Crucially, no Janus effect was detected, meaning the direction of the association did not reverse across any specification.

These findings stand in contrast to the substantial proportion of the literature reporting negative associations between screen time and well-being (Twenge & Campbell, 2018). They are, however, broadly consistent with the null or attenuated findings observed in EMA-based studies using objective usage measures (Bradley & Howard, 2023) and with the small effect sizes documented by Orben & Przybylski (2019) in large-scale survey data. The positive direction found here aligns with Minich & Moreno (2024), who similarly found that real-world smartphone use was associated with improved rather than worsened mood among adolescents.

Rather than reporting an isolated finding tied to arbitrary analytical paths, the VoE framework provides a transparent evaluation of how robust the overall conclusion remains across different modeling choices (Patel et al., 2015). In the present case, the choice of covariate set was the most influential decision node: models adjusting for demographic variables—either on their own or combined with conscientiousness—produced the largest effects, while models adjusting only for neuroticism or sex produced the smallest. This sensitivity to covariate selection is theoretically interpretable. Neuroticism is known to amplify affective reactivity and to predict both problematic smartphone use and lower well-being (Hisler et al., 2020; Peterka-Bonetta et al., 2019), so its inclusion as a covariate plausibly absorbs variance otherwise attributed to phone use. By contrast, adjusting for demographic group removes age-related confounds in both variables, revealing a slightly stronger residual association.

Predictor coding was the second-most consequential choice. The log transformation consistently yielded larger effect sizes than raw hours, suggesting that the association is somewhat more detectable when the influence of heavy users is attenuated. This observation carries methodological implications: researchers who do not report both codings may systematically understate or overstate the association depending on their operationalization choice. Model weighting had a

negligible impact throughout, indicating that participants with more EMA observations were not systematically different in the strength of their smartphone use-mood relationship.

Several limitations should be noted. The analysis relied on person-level aggregates rather than the full EMA time series, which means that within-person dynamics could not be examined. The results therefore speak to between-person robustness only. The sample consisted exclusively of Italian university students, a demographic that limits generalizability to other age groups, cultural contexts, and clinical populations. Furthermore, smartphone use was operationalized as total active app use, without distinguishing between passive scrolling, active communication, or functional activities, a distinction that the broader literature suggests may matter considerably for well-being outcomes (Verduyn et al., 2017). Perhaps most fundamentally, no predefined specification space is fully objective: the analytical decisions included in the multiverse inevitably reflect the researcher’s theoretical assumptions and judgment (Del Giudice & Gangestad, 2021), and a different researcher might have constructed a different set of specifications and reached partially different conclusions. This is not a weakness unique to VoE, but a reminder that transparency about how the specification space was built is as important as the analysis itself.

4.9 Conclusions

This chapter applied VoE analysis to EMA data on smartphone use and mood, estimating 100 model specifications that varied in predictor coding, covariate set, and model weighting. The results converge on a clear pattern: the association between mean daily smartphone app use and mean mood is consistently positive in direction, small in magnitude, and statistically non-significant across all specifications. No Janus effect was detected.

The principal contribution of this analysis is not a definitive estimate of the smartphone use-mood association, but a principled demonstration of its analytic stability. The direction of the effect does not depend on modeling choices; its magnitude varies modestly and predictably as a function of covariate selection. These findings illustrate how VoE analysis can inform interpretation while simultaneously fulfilling the methodological goal of transparency in the face of analytic flexibility.

Chapter 5

Conclusions

This thesis aimed at presenting the *Vibration of Effects* (VoE) framework as a methodological tool for assessing the robustness of empirical findings in psychological research. The replication crisis served as the foundational context: a substantial proportion of published psychological findings fail to reproduce and replicate under independent investigation (Open Science Collaboration, 2015), raising serious concerns about the credibility of the discipline. This crisis is driven by structural factors such as *Questionable Research Practices* (QRPs), publication bias, and the inflation of Type I error rates arising from *researcher degrees of freedom* (RDFs; Simmons et al. (2011)). Practices such as *p*-hacking, HARKing, and selective reporting systematically distort the published record. The theory crisis (Eronen & Bringmann, 2021) and the validation crisis (Schimmack, 2021) were identified as deeper expressions of the same problem. Methodological innovations such as preregistration, open data, and multiverse approaches were introduced as promising responses, with a focus on VoE.

The second chapter examined the role of effect sizes in psychological inference. A *p*-value indicates whether observed data are compatible with the null hypothesis, but it does not describe how large an effect is (Kelley & Preacher, 2012). Standardized measures such as Cohen’s *d* and Pearson’s *r* address this by quantifying effect magnitude on a scale-free metric (Cohen, 1988), yet their interpretation cannot rest on fixed numerical benchmarks alone. An effect’s practical value depends on context: a small effect that repeats daily, such as a minor personality trait or a subtle social influence, can accumulate into substantial differences in health, social status, and

overall well-being. An effect that appears negligible at the individual level may carry profound consequences when scaled across large populations (Funder & Ozer, 2019). Evaluating estimates through context-sensitive tools such as the *critical effect size* (CES) allows researchers to move beyond binary significance toward a more consequence-based approach to inference (Perugini et al., 2025). At the same time, in underpowered studies, effect size estimates are prone to systematic distortion: *Type M errors* inflate the observed magnitude, while *Type S errors* reverse the direction of the effect (Gelman & Carlin, 2014). These concerns motivate VoE analysis, where examining the distribution of estimates across model specifications provides a more complete and transparent picture of the data than any single estimate can offer.

The third chapter presented the theoretical and methodological foundations of VoE. It began with an introduction to the *garden of forking paths* (Gelman & Loken, 2013), which captures the idea that each analytical decision is a branching point that can lead to different conclusions from the same data. *Multiverse Analysis* (MA; Steegen et al. (2016)) and VoE (Patel et al., 2015) were introduced as tools that address this problem by estimating effects across the full range of defensible analytical choices. While VoE represents the primary focus of this thesis, inferential approaches such as *Specification Curve Analysis* (SCA; Simonsohn et al. (2020)) and Post-selection Inference in *Multiverse Analysis* (PIMA; Girardi et al. (2024)) were also briefly discussed to provide a complete picture of the multiverse landscape, as they extend exploratory frameworks by testing formal hypotheses across the predefined specification space. The chapter then described the main visualization tool of VoE, the *vibration plot*, which plots each specification’s effect size against its *p*-value. Drawing on Patel et al. (2015) original study, two summary indices were presented: the *Relative Hazard Ratio* (RHR), which captures how much effect magnitude varies across specifications, and the *Relative p-value* (RP), which tracks fluctuations in statistical significance. Both indices were described alongside their graphical representations, illustrating how the distribution of estimates across the predefined specification space can be read and interpreted visually. The chapter then turned to two case studies, on α -tocopherol and environmental cadmium, to illustrate the two main patterns that can emerge from a VoE analysis: the *Janus effect*, where the direction of the association flips depending on model choice, and vibration of magnitude, where direction is stable but effect size varies substantially. The chapter concluded by discussing the limitations of VoE, including the risk of specification noise, computational demands, and the difficulty of translating distributional findings into clear practical recommendations (Del Giudice & Gangestad, 2021).

The fourth chapter applied the VoE framework to an *Ecological Momentary Assessment* (EMA) dataset collected at the University of Trento ($N = 126$) as part of the WeNet project. The chapter began by describing the study design and the data collection procedure, including the methodological choice of keeping the analytic sample constant across the specification space, so as to effectively separate the influence of modeling choices from the influence of sample composition. 100 model specifications were estimated, varying systematically across three analytical decision nodes: predictor coding (raw vs. log-transformed hours of use), covariate set (25 theoretically grounded options), and model weighting (unweighted vs. weighted by number of valid EMA responses). Results were presented through a series of visualizations, including: a distribution of effect sizes across specifications, a volcano plot mapping each specification’s partial correlation against its statistical significance, and a specification curve ordering all estimates from smallest to largest alongside the analytical choices that produced them.

Importantly, the primary contribution of this thesis is methodological rather than substantive. The smartphone use-mood association was selected as a realistic psychological case study through which the VoE framework could be illustrated and evaluated. By applying VoE to real-world psychological data, this work demonstrates how researchers can systematically examine the robustness of effect size estimates and identify the extent to which conclusions depend on alternative analytical decisions. Across all 100 specifications, the partial correlation was consistently positive, ranging from $r = .034$ to $r = .097$ ($Mdn = .077$). No specification reached statistical significance ($p < .05$), and no Janus effect was detected. The covariate set proved to be the most influential decision node: models adjusting for demographic variables produced the largest effects, while models controlling only for neuroticism or sex produced the smallest. This pattern is theoretically interpretable, as neuroticism is associated with both greater smartphone use and lower well-being (Hisler et al., 2020; Peterka-Bonetta et al., 2019), and its inclusion absorbs variance that would otherwise be attributed to phone use. Predictor coding had a moderate influence, with the log-transformed predictor yielding larger estimates, while model weighting made little difference throughout. These findings are broadly consistent with EMA studies using objective usage measures (Bradley & Howard, 2023) and with the small effects reported in large-scale survey research (Orben & Przybylski, 2019), and the positive direction aligns with Minich & Moreno (2024), who found that real-world smartphone use was linked to better, not worse, mood in adolescents. An important lesson emerging from this work is that statistical significance and robustness should not be treated as equivalent concepts.

An effect may be statistically significant in some specifications yet highly sensitive to analytical decisions, whereas a non-significant effect may remain remarkably stable across alternative analyses. The VoE framework helps make this distinction explicit by shifting attention from isolated results to the overall distribution of effect size estimates. In doing so, it encourages a more nuanced evaluation of empirical evidence based not only on whether an effect is detected, but also on how consistently it emerges across defensible analytical choices.

The present work carries several limitations. The analysis relied on person-level aggregates rather than the full EMA time series, which means that within-person dynamics could not be examined. The results therefore speak to between-person robustness only. The sample consisted exclusively of Italian university students, a demographic that limits generalizability to other age groups, cultural contexts, and clinical populations. Furthermore, smartphone use was operationalized as total active app use, without distinguishing between passive scrolling, active communication, or functional activities, a distinction that the broader literature suggests may matter considerably for well-being outcomes (Verduyn et al., 2017). Perhaps most fundamentally, no predefined specification space is fully objective: the analytical decisions included in the multiverse inevitably reflect the researcher’s theoretical assumptions and judgment (Del Giudice & Gangestad, 2021), and a different researcher might have constructed a different set of specifications and reached partially different conclusions. This is not a weakness unique to VoE, but a reminder that transparency about how the specification space was built is as important as the analysis itself.

Future research should address these gaps. Applying VoE directly to the EMA time series using multilevel models would allow within-person robustness to be examined, exploiting the temporal richness that motivates EMA designs (Trull & Ebner-Priemer, 2013). Distinguishing between types of smartphone use, such as social media, communication, entertainment, and productivity, would help clarify whether the association with mood depends on how, rather than how much, devices are used.

The central finding of this thesis is that, across 100 model specifications, the association between smartphone use and mood was consistently positive, small, and non-significant, with no Janus effect emerging under any analytical pathway. This result illustrates both the strengths and the boundaries of VoE analysis. It can demonstrate that a finding is directionally stable, showing that no particular analytical choice is required to obtain a positive association. It cannot, however, establish that the effect is practically meaningful or causally valid: a small, non-significant association

remains small and non-significant regardless of how robust it is. At a broader level, this thesis has argued that robustness analysis is not an optional addition to standard reporting but a foundational requirement for interpretable inference. When analytical decisions go unreported, published effect sizes reflect one arbitrary path through the data rather than a stable empirical property of the phenomenon under study (Gelman & Loken, 2013). VoE, by making the predefined specification space explicit and transparent, proves to be a concrete and practicable methodological tool that directly supports the principles of open science, promoting reproducibility, analytical transparency, and greater credibility in psychological research.

More broadly, the present thesis illustrates how robustness analyses can contribute to addressing some of the methodological challenges highlighted by the replication crisis. By explicitly examining the sensitivity of effect size estimates to alternative analytical decisions, VoE helps researchers identify findings that depend heavily on certain specifications before they enter the published literature. In this sense, robustness analyses may represent a valuable complement to replication efforts, promoting greater transparency and more realistic expectations regarding the stability of empirical results.

References

- Aczel, B., Szaszi, B., Nilsson, G., Akker, O. R. van den, Albers, C. J., Assen, M. A. van, Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., Dongen, N. N. van, Donkin, C., Doorn, J. B. van, ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, 10, e72185. <https://doi.org/10.7554/eLife.72185>
- Altoè, G. (2023, September 19). *Replicability crisis is not the only one: Other current challenges/opportunities to enhance science*. Workshop: Replicability crisis in science?, Padova, Italy. Paper presentation.
- Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagni, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.02893>
- American Psychological Association. (2018). *Reliability*. American Psychological Association. <https://dictionary.apa.org/reliability>
- Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543-554. <https://doi.org/10.1177/1745691612459060>
- 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. *Journal of Management*, 42(1), 520. <https://doi.org/10.1177/0149206315619011>
- Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391), 531-533. <https://doi.org/10.1038/483531a>

- B.Nezlek, J. (2011). *Multilevel Modeling for Social and Personality Psychology*. SAGE Publications Ltd. <https://doi.org/10.4135/9781446287996>
- Borenstein. (2009). *Introduction to meta-analysis* (p. ixix). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470743386.fmatter>
- Bradley, A. H. M., & Howard, A. L. (2023). Stress and Mood Associations With Smartphone Use in University Students: A 12-Week Longitudinal Study. *Clinical Psychological Science*, 11(5), 921–941. <https://doi.org/10.1177/21677026221116889>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmeld, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436. <https://doi.org/10.1126/science.aaf0918>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, 6(1), 2942. <https://doi.org/10.1038/s41562-021-01193-7>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145153. <https://doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, 175(9), 526–536. <https://doi.org/10.1097/00005053-198709000-00004>
- Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6(3), 281–294. <https://doi.org/10.1007/BF02138940>
- Cumming, G. (2013). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis* (1st ed.). Routledge. <https://doi.org/10.4324/9780203807002>
- Del Giudice, M., & Gangestad, S. W. (2021). A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920954925>

- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511761676>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779788. <https://doi.org/10.1177/1745691620970586>
- Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE*, 5(4), e10271. <https://doi.org/10.1371/journal.pone.0010271>
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1), 4552. <https://doi.org/10.1177/1948550615612150>
- 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), 456465. <https://doi.org/10.1177/2515245920952393>
- Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association : JMLA*, 105(2), 203–206. <https://doi.org/10.5195/jmla.2017.88>
- 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), 156168. <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>
- Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 117.
- Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagni, A., & Finos, L. (2024). Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test. *Psychometrika*, 89(2), 542568. <https://doi.org/10.1007/s11336-024-09973-6>
- Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the netherlands. *PLOS ONE*, 17(2), 116. <https://doi.org/10.1371/journal.pone.0263023>

- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. <https://doi.org/10.1007/s10654-016-0149-3>
- 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>
- Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis = trust crisis? The effect of successful vs failed replications on laypeople’s trust in researchers and research. *Public Understanding of Science (Bristol, England)*, 29(3), 270288. <https://doi.org/10.1177/0963662520902383>
- Hisler, G., Twenge, J. M., & Krizan, Z. (2020). Associations between screen time and short sleep duration among adolescents varies by media type: evidence from a cohort study. *Sleep Medicine*, 66, 92–102. <https://doi.org/10.1016/j.sleep.2019.08.007>
- 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>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524532. <https://doi.org/10.1177/0956797611430953>
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). Guilford Press.
- Kathawalla, U.-K., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, 7(1), 18684. <https://doi.org/10.1525/collabra.18684>
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2), 137–152. <https://doi.org/10.1037/a0028086>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196217. https://doi.org/10.1207/s15327957pspr0203_4
- Klau, S., Schönbrodt, F. D., Patel, C. J., Ioannidis, J. P. A., Boulesteix, A.-L., & Hoffmann, S. (2023). Comparing the vibration of effects due to model, data pre-processing and sampling

- uncertainty on a large data set in personality psychology. *Meta-Psychology*, 7. <https://doi.org/10.15626/MP.2020.2556>
- 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), 142152. <https://doi.org/10.1027/1864-9335/a000178>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Psychological Science*, 62(3), 221230. https://doi.org/10.24602/sjpr.62.3_221
- Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science*, 16(3), 639–648. <https://doi.org/10.1177/1745691620958012>
- Lewandowsky, S., & Bishop, D. (2016). Research integrity: Don’t let transparency damage science. *Nature*, 529(7587), 459–461. <https://doi.org/10.1038/529459a>
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584585. <https://doi.org/10.1126/science.aal3618>
- Machery, E. (2020). What is a replication? *Philosophy of Science*, 87(4), 545567. <https://doi.org/10.1086/709701>
- Mayo-Wilson, E., Li, T., Fusco, N., Bertizzolo, L., Canner, J. K., Cowley, T., Doshi, P., Ehmsen, J., Gresham, G., Guo, N., Haythornthwaite, J. A., Heyward, J., Hong, H., Pham, D., Payne, J. L., Rosman, L., Stuart, E. A., Suarez-Cuervo, C., Tolbert, E., ... Dickersin, K. (2017). Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of Clinical Epidemiology*, 91, 95–110. <https://doi.org/10.1016/j.jclinepi.2017.07.014>
- Minich, M., & Moreno, M. (2024). Real-world adolescent smartphone use is associated with improvements in mood: An ecological momentary assessment study. *PLOS ONE*, 19(5), e0298422. <https://doi.org/10.1371/journal.pone.0298422>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-017-0021-0>

016-0021

- Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*, 34(1), 1–17. <https://doi.org/10.1007/s10869-017-9524-7>
- National Academies of Sciences, E., & Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. <https://doi.org/10.17226/25303>
- 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., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 14221425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 26002606. <https://doi.org/10.1073/pnas.1708274114>
- 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., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(Volume 73, 2022), 719748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 15961618. <https://doi.org/10.3758/s13423-019-01645-2>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 3(2), 173–182. <https://doi.org/10.1038/s41562-018-0506-1>
- Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 5(11), 1535–1547. <https://doi.org/10.1038/s41562-021-1141-1>

- Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528530. <https://doi.org/10.1177/1745691612465253>
- Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 10461058. <https://doi.org/10.1016/j.jclinepi.2015.05.029>
- Patil, P., Peng, R. D., & Leek, J. T. (2016). A statistical definition for reproducibility and replicability. *bioRxiv : The Preprint Server for Biology*. <https://doi.org/10.1101/066803>
- Perugini, A., Toffalini, E., Gambarota, F., Lakens, D., Pastore, M., Finos, L., & Altoè, G. (2025). *The benefits of reporting critical effect size values*. <https://doi.org/10.31234/osf.io/7qe92>
- Peterka-Bonetta, J., Sindermann, C., Elhai, J. D., & Montag, C. (2019). Personality Associations With Smartphone and Internet Use Disorder: A Comparison Study Including Links to Impulsivity and Social Anxiety. *Frontiers in Public Health*, 7, 127. <https://doi.org/10.3389/fpubh.2019.00127>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74(2), 166–169. <https://doi.org/10.1037/0022-0663.74.2.166>
- Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13(1), 19–30. <https://doi.org/10.1037/1082-989X.13.1.19>
- Scheel, A. M., Schijen, M. R. M. J., & 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. <https://doi.org/10.1177/25152459211007467>
- Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, 5. <https://doi.org/10.15626/MP.2019.1645>
- 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), 90100. <https://doi.org/10.1037/a0015108>
- Sewall, C. J. R., & Parry, D. A. (2024). Social media and youth mental health: Simple narra-

- tives produce biased interpretations. *Journal of Psychopathology and Clinical Science*, 133(7), 507–514. <https://doi.org/10.1037/abn0000950>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69, 487–510. <https://doi.org/10.1146/annurev-psych-122216-011845>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Steen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3), 279–282. <https://doi.org/10.4300/JGME-D-12-00156.1>
- Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M., Giangregorio, L., Dennis, B., Kosa, D., Debono, V. B., Dillenburg, R., Fruci, V., Bawor, M., Lee, J., Wells, G., & Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology*, 13(1), 92. <https://doi.org/10.1186/1471-2288-13-92>
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9, 151–176. <https://doi.org/10.1146/annurev-clinpsy-050212-185510>
- Twenge, J. M., & Campbell, W. K. (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive Medicine Reports*, 12, 271–283. <https://doi.org/10.1016/j.pmedr.2018.10.003>
- Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S. P. (2011). Replication in Prevention Science. *Prevention Science*,

- 12(2), 103–117. <https://doi.org/10.1007/s11121-011-0217-6>
- 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. <https://doi.org/10.1177/09637214211067779>
- Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review: Do social network sites enhance or undermine subjective well-being? *Social Issues and Policy Review*, 11, 274–302. <https://doi.org/10.1111/sipr.12033>
- Vinatier, C., Hoffmann, S., Patel, C., DeVito, N. J., Cristea, I. A., Tierney, B., Ioannidis, J. P. A., & Naudet, F. (2025). What is the vibration of effects? *BMJ Evidence-Based Medicine*, 30(1), 61–65. <https://doi.org/10.1136/bmjebm-2023-112747>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. J. van der, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, Volume 7 - 2016*. <https://doi.org/10.3389/fpsyg.2016.01832>
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>
- Yang, Y. S., Ryu, G. W., & Choi, M. (2019). Methodological Strategies for Ecological Momentary Assessment to Evaluate Mood and Stress in Adult Patients Using Mobile Phones: Systematic Review. *JMIR mHealth and uHealth*, 7(4), e11215. <https://doi.org/10.2196/11215>