The Ideal Psychologist vs. a Messy Reality: Using and Misunderstanding Effect Sizes, Confidence Intervals and Power

Elizabeth Collins

A thesis submitted for the degree of Doctor of Philosophy

University of Stirling
Division of Psychology
First Submission Date: January 2022

Thesis Abstract

Background: In the past two decades, there have been calls for statistical reform in psychology. Three key concepts within reform are effect sizes, confidence intervals and statistical power. The aim of this thesis was to examine the use and knowledge of these particular concepts, to examine whether researchers are suitably equipped to incorporate them into their research.

Methods: This thesis consists of five studies. Study 1 reviewed author guidelines across 100 psychology journals, to look for any statistical recommendations. Study 2 (*n = 247*) and Study 3 (*n = 56*) examined the use and knowledge of effect sizes using a questionnaire and online experiment. Study 4 surveyed psychology researchers on their use and knowledge of confidence intervals (*n = 206*). Similarly, Study 5 surveyed psychology researchers on their use and knowledge of power analyses and statistical power (*n = 214*).

Findings and Conclusions: Typically, psychology journals expect authors to report effect sizes in their work, although there are fewer expectations related to confidence intervals. Power analyses are also frequently encouraged for sample size justification. Self-reported use of effect sizes, confidence intervals and power analyses was high, while common barriers to use included a lack of knowledge, a lack of motivation, and the influence of academic peers. While knowledge of effect sizes was quite high, they appear to only be understood in relatively limited contexts. In contrast, both confidence intervals and statistical power appear to be frequently misunderstood, and many researchers find power analysis calculations difficult. Researchers would benefit from increased education and support to encourage them to confidently adopt an assortment of statistics in their work, and more effort must be made to prevent statistical changes from becoming a new series of tick-box exercises that do not improve the integrity of psychological research.

Acknowledgements

I am incredibly fortunate to have so many people to thank for supporting me through this PhD journey. Thank you first to the ESRC, who believed in my proposal and funded this research, and thank you also to my team of fantastic supervisors: to Line Caes for your careful proof reading and support, and to Eoin O'Sullivan for your contagious enthusiasm. Of course, my biggest thank you must go to my primary supervisor, Roger Watt. Thank you for being perhaps the most unconventional, yet exceptional, supervisor that anyone could ever hope for. It has been a joy and an honour to work with you.

I am lucky enough to have found my tribe several times over while working on this PhD, despite the limits that the pandemic has enforced on us all. In my department I have been surrounded by strong and successful women, and could not have done this without you all: Becca, Sarah, Sophia, Jordan, Gemma, Juliet, Kirsten, Danielle, Hannah and Louise. To Jordan especially – thank you for planting a virtual forest with me over the past year to keep me focused, for all your proof reading support, and for your endless positivity. Outside of my department I also have to thank Maddi, Jenny, the rest of the RoSE group, my PsyPAG colleagues, my PhD Women Scotland blog team, the Nowhere Lab, and so many others for the friendship, kindness and solidarity I have been lucky enough to experience on this journey.

Thank you of course to my family, who don't really understand academia at all but have smiled and asked questions anyway, and who are always proud. And thank you to my in-laws, who do very much understand PhDs, and have similarly never stopped supporting me. This PhD was also made easier by being surrounded by friends who are out there in the 'real world', to keep everything in perspective. Thank you especially to Lauren, Laura, Emma and Luke for still managing to be interested in my work after so many years of doing what seems like the same thing, and for cheering me on as I head to the finish line.

Finally, thank you to my partner, Tristan. Words cannot express how grateful I am for all your support over the past few years. You have believed in me every single step of the way, even when I wasn't so sure of myself (and even though you don't have a clue what this thesis is about). I could never have achieved this without you, and I'm *so* excited for whatever adventure is next.

Publications and Conference Presentations Associated With This Thesis

Collins, E., & Watt, R. (2021). Using and Understanding Power in Psychological Research: A Survey Study. *Collabra: Psychology*, *7*(1), 28250. https://doi.org/10.1525/collabra.28250

Collins, E. (2021, July 29-30). Using Power Analysis in Psychology. PsyPAG Annual Conference. Online.

Collins, E. (2020, July 31). Effect Size Use in Psychology. PsyPAG Annual Conference. Online. https://osf.io/tsz2c/

COVID-19 Impact: Due to the coronavirus pandemic, many conferences were cancelled in 2020, and subsequently pivoted to smaller online events in 2021. Several also narrowed the scope of submissions to focus on online research and the pandemic. This negatively impacted the dissemination opportunities for my work. However, I was fortunate enough to present twice at PsyPAG, in addition to giving smaller presentations within my university department.

Declaration

I declare that, except where explicit reference is made to the contribution of others, the thesis embodies the results of my own research and was composed by me.

This thesis has not been submitted for any other degree at the University of Stirling or any other institution.

Signature: 

Printed name: Elizabeth Collins

Date of first submission: 28/01/2022

Contents

List of Tables

List of Figures

Chapter 1: Thesis Introduction

Typical quantitative psychological research involves collecting data from samples, and using this data to make inferences about the population of interest. Statistics bridge the gap between samples and populations, with the most popular inferential approach being null hypothesis significance testing (NHST). Briefly, NHST is a statistical analysis procedure which computes a *p*-value, a number which indicates the probability of the found effect (or greater) occurring by chance under the null hypothesis. If the *p*-value is below a particular threshold (often 0.05), a result is categorised as statistically significant and the null hypothesis is rejected.

The use of NHST within psychology has become problematic for a number of reasons, but primarily because journals have historically been biased towards publishing statistically significant findings. Years of criticism of this publication bias, coupled with wider statistical issues and the recent replication crisis, have brought about a period of statistical reform in psychology. Many individuals and organisations now call for NHST to be accompanied or replaced by alternative statistical methods, including Bayesian analyses (e.g. Wagenmakers et al., 2018), or the estimation approach (e.g. Calin-Jageman and Cumming, 2019), which prioritises the use of effect sizes and confidence intervals for reporting and evaluating data. Others who focus instead on *improving* the use of NHST propose alternatives such as banning the concept of a threshold for statistical significance (e.g. McShane et al., 2019), adopting modified 'second generation *p*-values' (e.g. Blume et al., 2019), or optimising the use of NHST through increased statistical power (e.g. Cohen, 1992; Maxwell, 2004).

This chapter first discusses the use and misuse of NHST and its connection to the replication crisis and to academic publishing culture, before focusing specifically on the estimation approach and statistical power as key concepts within statistical reform. This focus will be justified with connections to the wider literature and the recommendations for statistical use shared by organisations such as the American Psychological Association (APA), and is further justified by the review presented in Chapter 2.

1.1 A Brief Overview of the Replication Crisis

In 2005, the paper '*Why Most Published Research Findings are False*' was published, which made the controversial claim that "*for most study designs and settings, it is more likely for a research claim to be false than true*" (Ioannidis, 2005, p. 124). Ioannidis' modelling examined the concept of the 'positive predictive value' (the post-study probability that a statistically significant study result is true, PPV), and estimated that it is less than 50% for most research. His suggested explanations for this low figure include the predominant use of small sample sizes, the desire for statistically significant *p*-values, and wider issues of researcher bias and unchecked flexibility in research design and analysis; much of which can be attributed to publication bias (discussed in Section 1.1.1). His modelling approach and claims have subsequently been criticised, as has the entirely hypothetical nature of the PPV (e.g. Goodman & Greenland, 2007; Jager & Leek, 2014; Morey, 2018). However, the broader ideas within Ioannidis' work contributed heavily to the growing conversation around research integrity.

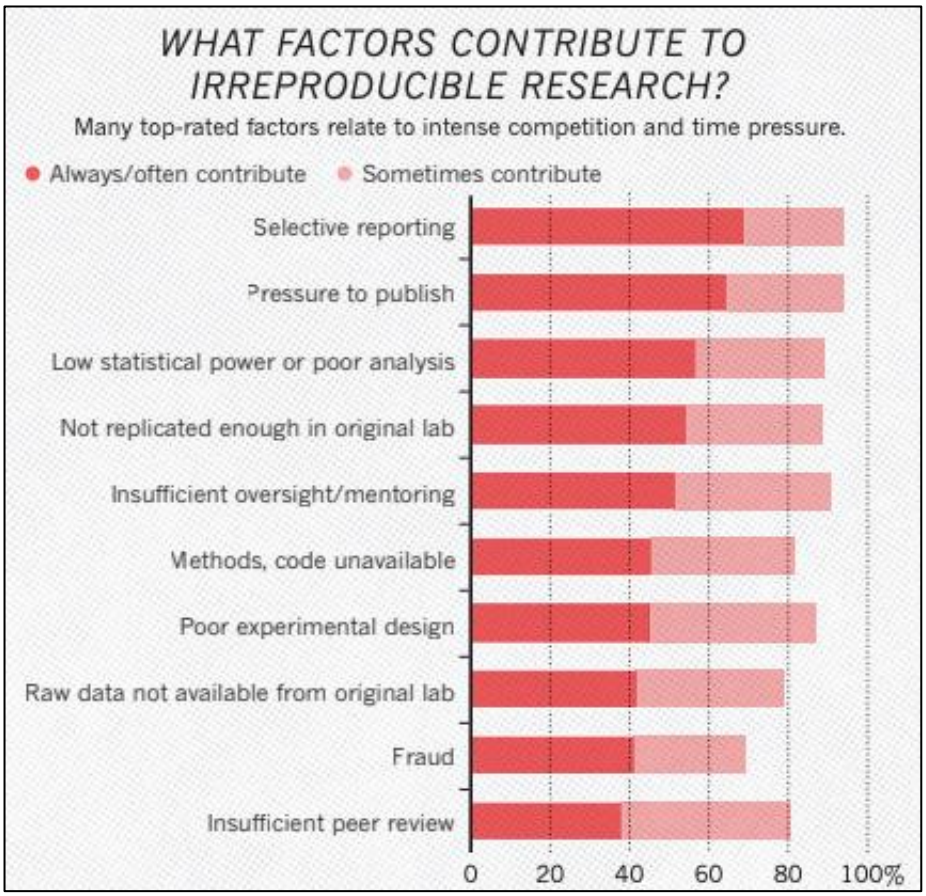
Several subsequent events and high-profile publications followed this paper, such as the discovery of long-running fraud by prominent researcher Diederik Stapel (Callaway, 2011) and the publication of a questionnaire where researchers admitting engaging in questionable research practices such as data manipulation and selective reporting (John et al., 2012). However, perhaps the pivotal moment of the replication crisis, and the moment which gave it its name, was the publication of the attempt to replicate 100 psychology studies with new, large samples of participants (Open Science Collaboration, 2015). Effect sizes in the replicated studies were, on average, half the size of those computed in the original studies, and just 36% of replications had statistically significant findings (versus 97% of originals). This negative perception of published psychology research was further compounded by a publication the following year, which identified a catalogue of errors across reported NHST findings (Nuijten et al., 2016), some of which directly impacted the conclusions made in the paper (e.g. mistakenly reporting a result as statistically significant). In light of the various controversies and events of the past two decades, the term 'replication crisis' has become a broader label used to capture issues of replication (repeating research on new samples), reproducibility (confirming results by analysing the same data set in the same way), and overall research integrity within psychology.

1.1.1 The Replication Crisis and Publishing Culture

Researchers have offered an assortment of explanations for the replication crisis, as shown in Figure 1.1 (Baker, 2016) and Figure 1.2 (Munafò et al., 2017). Perhaps the most serious explanation for issues is fraud, where results have been deliberately falsified; but other reasons include poor research design, insufficient mentoring (leading to incorrect decisions), and a lack of open data (meaning results cannot be verified). One particularly important reason presented in Figure 1.1 is 'pressure to publish'; which is an explicit reflection on the widely-criticised nature of academic publishing culture.

Figure 1.1

The Causes of Irreproducible Research, Shared by Researchers



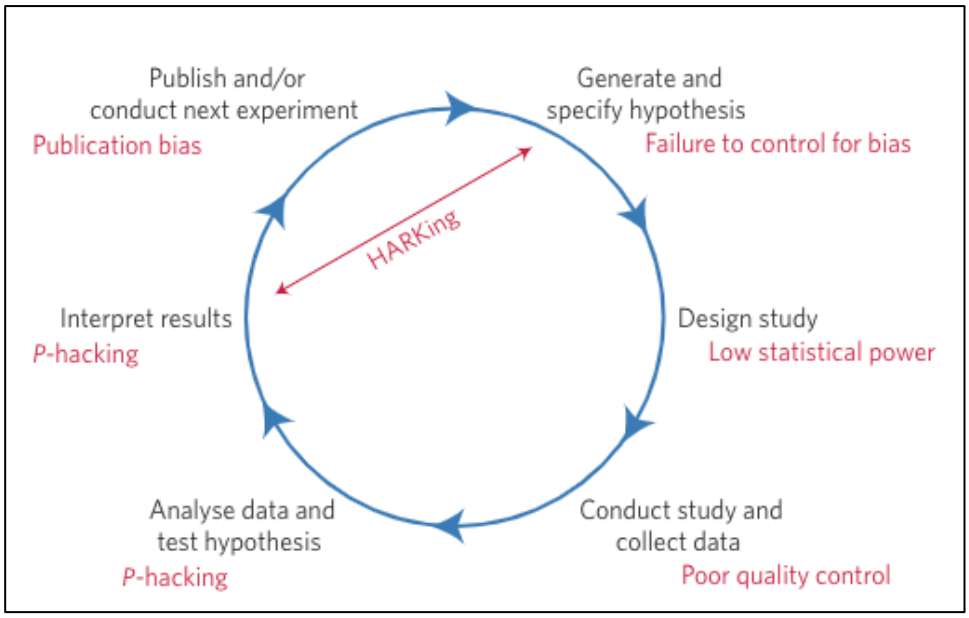
Note. From *1500 Scientists Lift the Lid on Reproducibility*, by M Baker, 2016, p. 453. Copyright 2016 by Macmillan Publishers Limited. Reprinted with permission.

Publishing papers in academia is described as "*the currency of academic science"* which *"increases the likelihood of employment, funding, promotion and tenure*" (Munafò et al., 2017, p. 7). It is essential for personal career progression, and is also the dominant form of

knowledge dissemination. Academic culture has the unofficial mantra of 'publish or perish' (e.g. Grimes et al., 2018), and quantity (of papers) has consistently been rewarded instead of quality. Several of the other explanations in Figures 1.1 and 1.2 can also partially be attributed to publication pressure, such as poor experimental design due to time pressure, and low statistical power due to small sample sizes (similarly attributable to time pressure). Low statistical power is a particularly prevalent criticism of psychological research, which will be addressed further in Section 1.4.

Figure 1.2

Threats to Reproducible Research



Note. From *A Manifesto for Reproducible Science*, by Munafò et al., 2017, p. 2. Copyright 2017 by Macmillan Publishers Limited. Shared under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

Beyond the publish or perish pressure, there is one crucial issue with academic publishing which has had a detrimental effect on research integrity: the prioritisation of novel, statistically significant results, and the frequent rejection of null findings (Ferguson & Heene, 2012). This bias towards significant findings is clear to see in reviews of the literature, which indicate that the proportion of psychology papers reporting statistically significant *p*-values sits at approximately 91.5% (Fanelli, 2010; Fanelli, 2012). As a consequence of publishing pressure and publication bias, researchers have historically discarded null findings (labelled the file-drawer problem; Rosenthal, 1979), and it is also claimed that many have employed questionable research practices to manipulate data to find a desirable significant result (Nosek

et al., 2012; Munafò et al., 2017), sometimes given the umbrella label of '*p*-hacking' (see Figure 1.2).

These questionable behaviours include rounding down *p*-values to meet the threshold for statistical significance, extending data collection to try and find data which is statistically significant, and excluding participants if it has a 'preferable' influence on the results. As noted above, many researchers have admitted to doing these things when surveyed anonymously (John et al., 2012). Horton describes this broadly negative culture within academia as a place where "*no-one is incentivized to be right*" (2015, p. 1380). Looking again at Figure 1.1 in more depth, several other explanations can arguably also be attributed to publication pressure and publication bias, such as selective reporting (to cherry-pick statistically significant findings), and indeed even outright fraud (to generate desirable findings to publish).

1.2 The Wider Issues with NHST

It is clear that the use of NHST, and publication bias towards statistically significant *p*-values, has had a detrimental effect on the reliability of psychological research. If NHST is simply problematic due to publication bias, then the solution appears to be simple: reform the publishing process to remove publication bias. The most popular proposed solution to this is to move to the registered reports model of peer review before data collection, so that papers are accepted based on their methodology instead of their findings. However, beyond institutional issues, the very nature of NHST is also problematic within many research contexts, which has led to wider calls for statistical reform.

First, an overview of the process of significance testing: mathematically speaking, NHST uses the sampling distribution for a null effect-size to produce a *p*-value, which ranges from 0-1. This *p*-value is the proportion of that distribution that exceeds the sample effect-size, and therefore is the probability that the found result, or a more extreme one, would occur should the null hypothesis be true. The calculated *p*-value is then assessed against a benchmark (typically .05) to determine whether or not a finding is statistically significant (where $< .05$ is significant). If the finding is declared statistically significant, the null hypothesis is rejected; and as discussed above, the result is much more likely to be accepted for publication.

Conceptually, NHST does not have one mathematical origin. Despite often mistakenly being attributed to Fisher, it is instead a jumbled confusion of Fisher's null hypothesis testing, and Neymar-Pearson's decision theory. It has borrowed (and distorted) Fisher's null hypothesis procedure, paired it with a binary decision inspired by Neymar and Pearson, and planted itself at the heart of research (Gigerenzer, 2004). This makes it particularly hard to interpret, and, as Cohen pointed out, it "*does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does*" (Cohen, 1994, p. 997). To put it clearly: *p*-values do not express the probability of a null hypothesis being true, nor the probability of an alternative hypothesis being untrue. In addition, they do not provide a measure of magnitude for the effect being studied. Furthermore, despite carrying a risk of errors (false positive or false negative findings*)*, *p*-values are typically presented and evaluated in a way which ignores any uncertainty, with researchers often erroneously claiming that their effect definitely exists because it is statistically significant (e.g. Hoekstra et al., 2006).

The spread of NHST has been attributed to popular texts such as the early textbook *Fundamental Statistics in Psychology and Education*, which (incorrectly) describes NHST as producing the probability that the null hypothesis is true (Guildford, 1942, as cited in Gigerenzer, 2004). Similarly, another popular 20th-century textbook authored by Nunally incorrectly described *p*-values as "*the probability that an observed difference is real*" and "*the degree of faith [that] can be placed in the reality of the finding*", alongside several other incorrect statements (Nunally, 1975, p. 194-195, as cited in Haller & Krauss, 2002). In line with increasing attention within academic textbooks, NHST use in published articles jumped from approximately 17% in the 1920s to over 90% in the 1970s and beyond (Hubbard & Ryan, 2000), and, as discussed, has become firmly embedded within publishing and research culture.

1.2.1 Misunderstanding NHST

In addition to, or perhaps as a result of, its lack of strong mathematical foundations, NHST and the resulting *p*-values are frequently misinterpreted by researchers. For instance, many researchers use the dichotomous significant-versus-not-significant judgement of a *p*-value as evidence of an effect either existing, or not existing (Meehl, 1978; Cohen, 1994; Hoekstra et al., 2006). Researchers also are guilty of claiming that statistical significance indicates that an

effect is important (Hoekstra et al., 2006); which is impossible to conclude without more information and context. Specific tests of NHST knowledge have demonstrated that *p*-value misunderstanding is widespread. For instance, in the late 20th century, Oakes (1986) presented a sample of 68 psychology researchers with statements such as "*you have found the probability of the null hypothesis being true*" and *"you can deduce the probability of the experimental hypothesis being true*", asking them to rate each statement as being either true or false (with all statements being false). He found that 97% of the sample incorrectly rated at least one of the statements as being true, with generally poor results overall. A subsequent replication by Haller & Krauss (2002) tested a further 69 psychology researchers, and found that approximately 85% of their sample also made at least one mistake about interpreting *p*-values, indicating very little improvement over time. Even more recently, these results still appear to be the case: for example, when replicated in a Spanish academic psychologist sample, 93.8% of participants made at least one mistake (Badenes-Ribera et al., 2015) and in a similar Chinese sample, nearly 100% of participants made at least one mistake (Lyu et al., 2018).

In addition to individual misunderstandings, reported *p*-values have also been shown to be rife with errors. Approximately half of psychology papers published between 1985 and 2013 which used NHST reported a *p*-value which did not match the associated test statistic or degrees of freedom (Nuijten et al., 2016). This review found that in 13% of the published articles which used NHST, mistakes were extreme enough to alter the statistical conclusions made by the researcher (e.g. a shift from statistical significance to non-significance). Whether this can be attributed to mistakes or fraud remains unknown, but nonetheless it further illustrates the poor reputation of NHST in psychology. It is clear that, publication bias aside, there are wider issues with NHST, and so reforming the publication process (e.g. through registered reports) is not necessarily a robust solution on its own.

1.2.2 A Future With or Without NHST

"*First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist."* (Cohen, 1994, p. 1001)

Opinions on how to remedy the issues surrounding NHST are varied. Some researchers propose that significance testing should be replaced by other methods, such as Bayesian

statistics (e.g. Rouder et al., 2009; Wagenmakers et al., 2018), a model comparison approach (e.g. Lavine, 2019), or an estimation approach focusing on effect sizes, confidence intervals and meta-analytic thinking (e.g. Calin-Jageman & Cumming, 2019). Others focus more strongly on modifying NHST, such as changing calculations to generate second generation *p*-values (Blume et al., 2019), using a stricter threshold for statistical significance, such as $p < .005$ instead of .05 (Benjamin et al., 2018) – or alternatively, removing the threshold altogether and retiring the use of the term 'statistically significant' (Amrhein et al., 2019; McShane et al., 2019). Further groups of researchers retain the belief that NHST should simply be better taught, and used more appropriately depending on the research design (e.g. Lakens, 2021). Others focus less on the specifics of statistical significance, and more on improving the overall reliability of NHST, through increased statistical power (and therefore larger sample sizes; e.g. Maxwell, 2004). These ideas represent just a small selection of proposals, and while each has their merits and their weaknesses, all embody some form of improvement beyond the current status quo.

Within all of the possibilities for statistical reform, this thesis focuses specifically on two directions listed above: the estimation approach (sometimes labelled the 'New Statistics'), and statistical power. These recommendations are central to those made at the organisational level, such as the APA. Within their Publication Manual, they call for interpretations of results to be based on calculated effect sizes and confidence intervals (APA, 2010, as cited in Cumming et al., 2012), and advises researchers to "*take statistical power seriously*" (APA, 2010, p. 30, as cited in Cumming et al., 2012, p. 143). These recommendations are further reinforced in their journal article reporting standards (JARS), which instruct researchers to "*describe the sample power*" and report any power analyses used, and include effect sizes and confidence intervals within the results sections of published works (Appelbaum et al., 2018). Furthermore, these statistics require a shift in mindset to consider uncertainty and estimation (in contrast to dichotomised NHST-thinking), without requiring a full transition to a new theoretical framework, such as Bayesian statistics. This is not intended to be a criticism of Bayesian statistics, but instead is an acknowledgement that the path of least resistance is likely to be the best attempt to engage researchers with statistical reform. Subsequent sections of this chapter will introduce the estimation approach and statistical power in more depth, before connecting these topics to the objectives of the thesis.

1.3 Statistical Solutions: The New Statistics

The estimation approach reflects a move away from the dichotomous and fixed approach to data evaluation, which is typically the result of using NHST, towards a more careful evaluation of findings based on the questions of 'how big is the effect?' and 'how certain is the finding?' (Cumming et al., 2012; Calin-Jageman & Cumming, 2019). This approach revolves around the use of effect sizes and confidence intervals, and additionally emphasises the use of meta-analyses to synthesise evidence for a given phenomenon. It has been given the nickname of the New Statistics to emphasise that they represent a modern change of perspective compared to the traditional or 'old' reliance on NHST (Cumming, 2012), although it is important to acknowledge that none of these statistical concepts are truly new. Note that within the estimation approach, this thesis focuses specifically on effect sizes and confidence intervals, because they are the most predominantly encouraged by the APA (as discussed in Section 1.2.2).

1.3.1 Effect Sizes

Put simply, an effect size is any value which quantifies the size of a phenomenon of interest: it is therefore a way to answer the question of 'how big is the effect?' when evaluating data. Something as simple as the difference between two means is an effect size: this falls into the category of *unstandardised* effect sizes, where the units of measurement are the same as those used in the study. The second type of effect size is the *standardised* type, which is considered 'unitless' and allows for more straightforward comparisons to be made with other effect sizes. The widely-used correlation index, Pearson's *r*, is an example of a standardised effect size, as is the equally common Cohen's *d*. Note that in this thesis, the term 'effect size' is used in the broadest sense to represent any kind of effect size index.

Calculating and reporting effect sizes provides multiple benefits. Principally, it presents a measure of magnitude, which is often more useful to a researcher than the information provided by a *p*-value. Effect sizes also encourage researchers to critically evaluate results generated using NHST, as a statistically significant finding may not necessarily translate to an effect size large enough to be of theoretical or practical interest. Furthermore, reporting effect sizes facilitates the computation of meta-analyses (combining estimates from multiple

studies), which are crucial for increasing the accuracy of knowledge about any particular phenomenon (Cumming, 2014).

Despite their apparent usefulness, effect sizes have been historically neglected in psychological research. While Pearson's *r* is a familiar effect size found across the historic literature, typically statistical analyses have not been accompanied by effect sizes beyond the basic output produced within software for each type of test. Cohen frequently promoted effect sizes throughout the 20th century, asserting that "*the primary product of a research inquiry is one or more measures of effect size, not p values*" (Cohen, 1965, as cited in Cohen, 1990, p. 1310). However, reviews of the literature in the mid-1990s found that effect sizes were present in just 10% of published papers (this figure includes *r* correlation values; Keselman et al., 1998). Even prior to the replication crisis, calls for their use grew; such as the APA stressing the importance of both reporting and interpreting effect sizes in their Task Force on Statistical Inference Report (Wilkinson, 1999). In more recent years, they have been incorporated into both the APA Publication Manual, and the more specific APA JARS guidelines; and subsequent reviews suggest there has been an upwards trend in effect size use. For instance, within a sample of articles published in 2002 and 2003 (i.e. the time period following the Task Force on Statistical Inference Report) effect sizes were reported in 62.5% of published articles (Dunleavy et al., 2006). A subsequent larger meta-review estimated that effect sizes are reported approximately 38% of the time in quantitative psychology research, with a clear pattern of increased use over time (Fritz et al., 2013).

1.3.2 Confidence Intervals

The second key element of the estimation approach is to ask 'how certain?'. Within the New Statistics, confidence intervals are promoted as the way to answer this question (although most recently, discussions under the New Statistics label have been extended to encourage the use of Bayesian credible intervals too; Calin-Jageman & Cumming, 2019). While effect sizes provide a measure of magnitude for sample data, confidence intervals provide inferences about the population of interest.

A confidence interval is a range of plausible population values, calculated using a sample estimate. Note that this can be a plausible range of population effect sizes, but is also commonly used to provide a range of plausible values for the population mean. They provide an explicit measure of uncertainty around a single estimate, providing more information than

a single value does alone; and are described by Cohen as *"surely a useful piece of knowledge in a science that presumes to be quantitative*" (1990, p. 1310). Accepting uncertainty and acknowledging it with statistics can generate several improved outcomes for psychology. Firstly, it encourages researchers to think meta-analytically, by bringing multiple studies together to consider all evidence, and by replicating findings. This provides more accurate point estimates and a more solid knowledge base. It also should encourage researchers to think more carefully about how each individual study is designed, such as minimising uncertainty by increasing sample sizes, and using effective approaches to measurement, generating research that is rigorous and reliable.

Confidence intervals appear far less frequently than effect sizes across the psychology literature. For example, several reviews of articles published in the late 20th century found zero evidence of confidence interval reporting (e.g. Keselman et al., 1998; Kieffer et al., 2001). In the same manner as effect sizes, the use of confidence intervals has also been strongly encouraged by the APA, having been included in both their Task Force report (Wilkinson, 1999) and both versions of the APA JARS guideline (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Appelbaum et al., 2018). However, despite promising increases in effect size use identified in Fritz et al.'s (2013) meta-review, the same cannot be said for confidence interval use. Their review estimated that confidence intervals are only included in 10% of published psychology papers, and use has not increased noticeably over time.

1.3.2.1 Confidence Interval Controversy

Despite their apparent usefulness, confidence intervals are more complex than they first appear, with strong disagreement regarding how they can and should be interpreted. Strictly speaking, a confidence interval is constructed and defined within a long-run probability, where one interval may or may not contain the true population value (Morey et al., 2016a). This means that 95% of (95%) confidence intervals calculated from repeated runs of the same experiment will do so, but 5% won't, and it is impossible to know whether one single interval does or does not. The common interpretation that a single confidence interval has a 95% chance of containing the true population value is mathematically incorrect (labelled the *fundamental confidence fallacy*; Morey et al., 2016), and some researchers conclude that the most straightforward way to interpret a confidence interval is to not interpret it at all. However, Miller and Ulrich challenge this perspective, by arguing that probability can still be

understood as a long-run perspective even within the context of a single event, using a deck of cards example. To quote their explanation (Miller & Ulrich, 2016, p. 127):

"Suppose that a standard deck of 52 playing cards was shuffled and placed on a table at 11:00 a.m., and then it was shuffled again at 11:05, with no one having examined it between shuffles. Then, at 11:05, you are asked whether it is true or false that "There is a 1/52 probability that the top card was the ace of spades at 11:01." If you are a strict frequentist, you must say this statement is false, because technically it refers specifically to the specific state of the deck at that time. As a matter of logic, the top card at 11:01 either was the ace of spades or it was not. If you had checked this card over and over - without reshuffling the deck, of course, since the statement concerns its state "at 11:01"- you would always have gotten the same result, so in the long run across these many redundant checks the probability would be 0 or 1.

Nevertheless, we think most people would say there was a 1/52 probability that the top card was the ace of spades at 11:01, because they do not think of a specific state of the deck as a single isolated outcome in the strict frequentist sense. Implicitly, they would interpret the probability statement as referring to a long-run scenario involving many shuffled decks of cards—not just to one specific isolated deck. When the question is viewed in terms of this long-run scenario, it is true that the ace of spades will be on top in 1/52 of the decks, and so it is correct to say that the probability of this event is 1/52 within this long-run scenario. This example illustrates that in common parlance the word "probability" is often taken to involve an implicit series of random events, especially when talking about random out-comes that are unknown — even if they have already been determined, like the identity of the top card in the shuffled deck."

The deck of cards example highlights a plausible way to interpret probability, although it is not one that is accepted by strict frequentist researchers such as Morey et al. (2016a; 2016b). In an attempt to move away from these disagreements over the mathematical understanding of probability, other researchers prefer to use more general language when interpreting confidence intervals. Instead of terms such as 'chance' or 'probability', several researchers argue that it is possible to be '95% *confident'* that a single interval contains the population value (e.g. Cumming, 2012; Garcia-Pérez & Alcalá-Quintana, 2016). Those who prefer the 95% confident approach argue that it is perfectly reasonable to understand an individual

confidence interval as being a plausible range of population values, provided that it is interpreted with the awareness that a small proportion of confidence intervals generated in the same way would not contain the true value (Cumming, 2012).

This debate on how to accurately interpret a confidence interval makes it clear that they may not be a straightforward replacement or supplement for NHST. This raises questions of how useful they could possibly be within statistical reform: if they are not easy to interpret, and cannot be agreed upon by researchers, how can the average researcher be expected to use them accurately? However, within the context of this thesis, confidence intervals have been chosen for inclusion because their use is being requested more frequently across psychology. In addition, they are traditionally presented as the second half of the New Statistics, to accompany effect sizes.

1.4 Statistical Solutions: Power

As discussed so far within this chapter, many options for statistical reform focus on moving away from NHST entirely, such as replacing or supporting *p*-values with statistics such as effect sizes and confidence intervals. These strategies focus on concerns regarding false positives (Type I errors) in the literature, as highlighted in Ioannidis' landmark paper (2005). In contrast, calls for increased statistical power are a method of *optimising* the use of NHST, as a way of reducing Type II errors (false negative results). As NHST remains prevalent within the literature, it is sensible to improve its use as much as is possible. Increased statistical power equals a reduced probability of a Type II error (false negative) result; and so, statistical power is a method of reducing the uncertainty of a set of results analysed using NHST. Statistical power was cited as a contribution to irreproducible research in Figure 1.1, and as noted in Section 1.2.2, is particularly central in organisational efforts to encourage statistical reform, such as the APA Publication Manual and APA JARS.

1.4.1 What is Power?

Statistical power (subsequently shortened to 'power' in this thesis) is the probability of obtaining a statistically significant outcome for a test, given a particular alpha level, sample size and population effect size. As sample sizes increase, so does power (when the other two factors remain the same). Power is tied to the Type II (false negative) error rate, where error

rates increase as power decreases, and the conventionally 'acceptable' level of power in psychological research is 80%, equal to a 20% chance of a Type II error (Cohen, 1992). While it would be logical to assume that researchers have long paid attention to power due to the historic prevalence of NHST, instead, psychological research has typically been underpowered due to inadequate sample sizes (e.g. Cohen, 1962; Stanley et al., 2018). Despite the cognitive disconnect between seeking significant *p*-values alongside inattention to power, there is one simple explanation: gathering larger samples takes more time and resources. In a publish or perish culture, academics have historically prioritised fast science, as explained by Vankov et al. (2014):

> *"Scientists are human and will therefore respond (consciously or unconsciously) to incentives; when personal success (e.g., promotion) is associated with the quality and (critically) the quantity of publications produced, it makes more sense to use finite resources to generate as many publications as possible" (Vankov et al., 2014, p. 1037).*

This disconnect raises questions about how the replication crisis can be partially blamed on small sample sizes and low power (e.g. Baker, 2016; Munafò et al., 2017), despite the abundance of statistically significant findings in the literature (Fanelli, 2010). If low power corresponds to more false *negatives*, i.e. null results, then how are there so many *positive* results in the literature? The most plausible answer is one that was explored in Section 1.1.1: questionable research practices, where data has been cherry-picked, manipulated and selectively reported to increase the chances of publication. While statistical power does not solve the issues of publication bias or data manipulation, nor does it have any impact on Type I errors, it does more broadly decrease the risk of Type II errors and therefore increase the reliability of NHST. As mentioned above, when sample sizes increase, so does power: hence, increasing sample sizes is the single best strategy to reduce the chance of Type II error when using NHST. Mathematically speaking, optimum sample sizes can be determined using *a priori* power analyses, as discussed below.

1.4.2 Power Analyses

As described above, power is the (1) probability of obtaining a statistically significant outcome for a test, given a (2) particular alpha level, (3) sample size and (4) population effect size. A *power analysis* is a calculation which uses three of these four pieces of information relevant to power to compute the fourth. Most frequently, this takes the form of an *a priori*

power analysis, which uses a fixed alpha, power level, and estimated population effect size to calculate a minimum sample size for a study with the chosen power level (contingent on the estimated population effect size being accurate). In most cases, when the broad term 'power analysis' is used, it is used to mean an *a priori* power analysis.

While these calculations are, at face value, rather straightforward, the 'population effect size' factor must be considered carefully. The obvious dilemma here is that in nearly all cases, the population effect size of any given phenomenon is unknown; and therefore power is a theoretical idea that cannot ever really be accurately computed. Instead, when we talk of power, we talk of a messier version of power, which relies on providing the best estimate of a population effect size, and using it in a power analysis calculation to get as close as possible to a well-powered study.

Two other types of power analysis also exist: post hoc (or 'observed') power analyses, and sensitivity power analyses. Historically, the more controversial post hoc power analysis has been a popular way to evaluate the power of a study. This takes the form of a retrospective calculation of a study's power based on the measured sample effect size, alpha and actual sample size. Post hoc power analyses have traditionally been used to suggest that null results are actually Type II errors, attributed to a lack of power (Onwuegbuzie & Leech, 2004). However, using the measured sample effect size is unlikely to be a reliable reflection of the true population effect size due to sampling error, meaning that it should not necessarily be used in a power calculation (Gelman, 2019). In addition, several researchers have demonstrated that post hoc power and *p*-values are directly related, and so when $p = .05$, post hoc power will be 50%, regardless of the combination of sample size and sample effect size, or of the true power of the study (Yuan & Maxwell, 2005; Lakens, 2014; Collins & Watt, 2021). Null results ($p > .05$) will always result in low post hoc power, regardless of the actual power of the study, rendering post hoc power analyses uninformative in most circumstances.

A more modern approach to power is a third type of power analysis: a sensitivity power analysis. Instead of estimating a population effect size, or (erroneously) relying on the sample effect size, a sensitivity power analysis uses the actual study sample size, alpha and power level to calculate the smallest sample effect size that a study could detect. This is often considered to be a more valuable reflective tool for evaluating the power of studies

(Bacchetti, 2010), as it does not require a researcher to produce an estimated effect size; nor rely on a biased estimate from a sample.

1.4.3 Power Beyond NHST

Mathematically speaking, statistical power is a frequentist concept which is interwoven with NHST. As NHST is criticised for failing to provide detailed information about many phenomena being studied, or for being misunderstood, the place of statistical power within the wider reform movement could also be criticised for perpetuating the use of NHST. However, it is highly unlikely that after decades of prevalence, NHST will simply disappear, and so improving its use is highly relevant as part of statistical reform. In addition, even when looking beyond NHST, it could be argued that statistical power still remains relevant in a broader way. If researchers adopt a mindset of increased sample sizes (whether that be with or without power analyses), they are naturally reducing sampling error and increasing the precision of the estimates they make, regardless of any theoretical or statistical framework they may rely on for data analysis.

1.5 Looking Forwards: Understanding Statistics

Thus far in this chapter, effect sizes, confidence intervals and statistical power have been introduced as three elements of statistical reform that may increase the transparency of reporting, reduce some of the uncertainty within NHST, and encourage more careful evaluation of results. However, it must also be asked: is simply reporting these statistics, and calculating power analyses, sufficient to generate successful statistical reform in psychology? While the estimation approach offers more detail for the careful evaluation of results, and reduced sampling error and Type II errors increase the reliability of sample data, one further issue that is not automatically addressed by using these processes is that of statistical *understanding*. NHST, with all its flaws, has been widely misunderstood by researchers, which undoubtedly contributes to its misuse (e.g. Haller & Krauss, 2002; Lyu et al., 2018); and so, alternatives are also at risk of becoming misunderstood and misused. Blindly adopting any kind of new statistical processes, particularly when most are automatically (and therefore often unthinkingly) calculated via software, may create a new crisis of mistakes which fail to increase the reliability of the field.

1.5.1 Understanding Effect Sizes and Confidence Intervals

Effect sizes sound like very simple and very broad concepts. However, at present, very little is known about how researchers understand them. Researchers need to be aware of the variety of effect size indices that exist, so that they are equipped with the knowledge to choose appropriate effect sizes for a given research context. In addition, while reporting effect sizes increases transparency and provides more information about a set of data, they primarily add value when they are used to evaluate the question of 'how big is this effect?'. This means that researchers need knowledge of not only what they are, but how they can be interpreted. Similarly, reporting confidence intervals provides an explicit measure of uncertainty; but their primary value is as an inferential statistic which provides evidence for the question of 'how certain?'; and so once again, researchers need sufficient knowledge to interpret them accurately. One further question raised by the use of either statistic could also be the extent to which researchers relate these 'new' concepts to NHST. Despite being unrelated, both may mistakenly be connected to NHST, increasing the likelihood that they may be misused or misunderstood by researchers who do not have a confident grasp on different statistics.

1.5.2 Understanding Power

With regards to power, it could be argued that a firm understanding is not essential, given that simply using a power analysis to plan increased sample sizes directly reduces sampling error and therefore increases the precision of a piece of research. However, if researchers do not really understand what statistical power is, they are less likely to either calculate or evaluate it accurately; and as historically, NHST has been misunderstood, it would be sensible to ensure that improvements to its use are not similarly misunderstood. In addition, effective power analyses rely on making sensible decisions about population effect sizes; which means that knowledge of effect sizes is also important.

1.6 Thesis Overview

This thesis examines the current state of statistical use and knowledge within the psychology researcher population, focusing specifically on the 'New Statistics' and statistical power due to their popularity within statistical reform efforts. In contrast to the current literature, which has typically offered broad reviews of statistical use within published articles, this thesis will

offer two alternative perspectives. First, statistical use will be studied through the lens of current journal expectations, given the influence that journals have on researcher behaviour, and then it will be studied through the individual experiences of researchers. While large organisations such as the APA strongly recommend that researchers adopt these statistics, little research currently exists to examine individual experiences, and consequently little is known about what may support or prevent researchers from changing their statistical behaviours.

While simply using effect sizes, confidence intervals and power analyses is a positive shift from relying solely on *p*-values, there is a risk that these statistical concepts will be misunderstood or misused in a similar manner as NHST, potentially rendering future research unreliable. Therefore, it is essential to also examine knowledge of each statistical concept within the psychology researcher population. This thesis will study knowledge of all three concepts discussed here: effect sizes, confidence intervals, and statistical power, using a combination of novel and replication materials.

1.6.1 Researcher Reflexive Statement

The research reported in this thesis was carried out in my position as an early career researcher, who has been exposed to peers and colleagues struggling with statistics for several years (throughout my student life, and through teaching and research roles). My first interest was in improved statistics training and resources for researchers, and so I co-authored a textbook on the subject (Watt and Collins, 2019). However, it became clear that more must be known about individual experiences and perspectives, along with knowledge levels, to more effectively educate and support researchers. My decision to focus on statistical power and the New Statistics was primarily due to their prevalence within the statistical reform efforts of psychology researchers. Personally, I believe that all three have value; although to differing extents. The question of 'how big is the effect?' is applicable to many research contexts, particularly when it is acknowledged that the term 'effect size' covers any kind of standardised or unstandardised measure. Similarly, the question of 'how certain?' is equally, if not more widely applicable. Confidence intervals, with all their mathematical limitations, are not the best answer to this, in my opinion. If the most intelligent researchers cannot agree on whether the use of '95% confident' is acceptable; and argue that one interval should not really be interpreted at all (Morey et al., 2016a), then it is difficult to encourage researchers with less statistical knowledge to use them without creating a risk of misunderstandings and

confusion. However, I believe that they are better than nothing, because encouraging researchers to acknowledge uncertainty is one small step towards improving research quality; not least because it encourages researchers to seek more and better evidence to decrease the uncertainty of their research. Finally, with regards to statistical power, I firmly believe that asking researchers to think more critically about sample size is a good thing; whether that is within or beyond the frequentist NHST framework. A further critical discussion of these statistics, their limitations, and their place within statistical reform is presented within Chapter 8 of this thesis, Section 8.4.

1.6.2 Thesis Objectives

This thesis has three key objectives:

1. To review the current contents of psychology journal author guidelines to identify the presence of any statistics guidelines, particularly looking for comments on NHST, or the inclusion of (1) effect sizes, (2) confidence intervals or (3) statistical power.

2. To examine how frequently psychology researchers report using each of the three aforementioned statistical concepts, along with their explanations for not using them.

3. To examine knowledge, understanding and interpretations of each statistical concept, identifying the prevalence of any misconceptions.

1.6.3 Thesis Structure

To achieve these three objectives, the studies reported in this thesis adopted a mixed-methods approach and are presented across seven research chapters, as detailed below.

Chapter 2 takes the form of a review of journal guidelines, establishing the current requirements for statistical reporting in psychological research. This chapter addresses Objective 1.

Chapter 3 presents the findings of an online questionnaire study, examining use and knowledge of effect sizes. This chapter addresses Objectives 2 & 3.

C*hapter* 4 presents an exploratory study, examining effect size estimation in the context of raw data shown on graphs. This chapter addresses Objective 3.

Chapter 5 presents the findings of a second online questionnaire study, exploring confidence interval use and knowledge. This chapter addresses Objectives 2 & 3.

Chapter 6 presents a second exploratory study, examining confidence interval interpretation using two research scenarios. This chapter addresses Objective 3.

Chapter 7 presents the findings of a third online questionnaire study, which explores the use and understanding of statistical power, a priori power analysis, and post hoc power analyses. This chapter addresses Objectives 2 & 3.

Chapter 8 presents the overall thesis discussion.

Chapter 2: A Review of the Statistical Guidelines of the Top 100 Psychology Journals

2.1 Abstract

Background: Journals are the dominant form of knowledge dissemination within academia, and academic researchers are often judged by the perceived prestige of the journals which publish their work. Author guidelines dictate the style and content of published articles, and it is argued that including statistical guidelines can increase the transparency, detail and quality of quantitative research. This chapter presents a review of the author guidelines of the top 100 psychology journals, providing details of any statistical guidelines that relate to the use of NHST, statistical power, effect sizes and/or confidence intervals.

Methods: The JCR Clarivate database was used to create a list of the top 100 psychology journals, excluding review-only and qualitative-only journals, as measured using the Clarivate impact factor. The author guidelines for each journal were then sourced and examined for specific statistical instructions relating to NHST, effect sizes, confidence intervals, and power/power analysis. Journal guidelines were also coded for references to organisational guidelines, overall accessibility, and the provision of statistical support resources, such as tutorials for authors.

Results: Twenty four of the top 100 psychology journals had no statistical requirements or recommendations for authors relating to NHST, effect sizes, confidence intervals, or statistical power. The remaining 76 had some form of requirements (*n = 26*), recommendations (*n = 40*) or mixed guidelines (*n = 10*). Overall, more journals requested effect size reporting compared to confidence intervals, and no journals asked that either be used *instead* of NHST. While more than half of journals asked authors to explain their sample size, fewer than half of journals encouraged or required authors to use power analyses to do so.

Conclusions: Statistical guidelines appear to be growing in prevalence across psychology journals, with less than a quarter of journals having no guidelines at all regarding statistics.

This is likely to have a long-term positive influence on reporting behaviour. However, as very few journals provide statistical resources for authors, there is an expectation on researchers (or reviewers) to ensure that they are using statistics correctly.

2.2 Introduction

As discussed in Chapter 1, psychology is in a period of transformation in the wake of the replication crisis, which called the reliability of psychological research into question. One factor which arguably has contributed to this crisis is psychology's reliance on null hypothesis significance testing (NHST), which dominates the published literature, textbooks, and university research methods classes. NHST is widely misunderstood and misinterpreted (e.g. Haller & Krauss, 2002; Lyu et al., 2018), often published with arithmetic mistakes (Nuijten et al., 2016), and is limited to providing information about the null hypothesis. This chapter briefly discusses how journals have contributed to the prevalence and misuse of NHST, considers the influence that journals can have on statistical reform in psychological research, and presents a review of the statistical guidelines of the top 100 psychology journals (as of October 2021).

2.2.1 The History of Journals and Statistics

The inferential analysis procedure of NHST continues to dominate psychological research, despite long-running criticisms (e.g. Gigerenzer, 2004). *P*-values can be found in more than 90% of published psychology articles (Cumming et al., 2007; Counsell & Harlow, 2017); and the majority of published *p*-values are statistically significant (Fanelli, 2010; Fanelli, 2012; Bakker et al., 2012). This imbalance is primarily attributed to publication bias, as introduced in Chapter 1. Publication bias is a consequence of a long-running aversion to null results, based on their difficulty to interpret (Ferguson & Heene, 2012), and the convenience of using statistical significance as a crude filter to keep accepted manuscripts at a manageable level. Indeed, in some instances, journal editors have even historically made it clear that they are looking for significant results to publish. One prime example of this is the editor of the Journal of Experimental Psychology from the 1960's, who insisted that researchers report *p*-values, and expressed a strong preference for submissions demonstrating $p < .01$ (Merton, 1962, as cited in Gigerenzer, 2004). And so over time, a statistically significant *p*-value has become a benchmark for publication, instead of a statistic that is used to support the evaluation of data in appropriate circumstances. As a consequence of this publication bias, researchers have historically discarded null findings (labelled the 'file-drawer problem'; Rosenthal, 1979), or even allegedly employed questionable research practices to manipulate data to find a desirable significant result (Nosek et al., 2012; John et al., 2012). Published

results are therefore a biased sample of the research that is done, are often unreliable, and arguably have contributed heavily to the replication crisis (Ioannidis, 2005).

2.2.2 Journals and Statistical Reform

Because journals are a dominant method of knowledge dissemination, and the majority of researchers must engage with the publishing process to communicate their research and advance their careers, journals can be a vital catalyst for statistical reform. Their historic priorities and decisions have strongly influenced statistical practices, and as such can arguably also influence statistical changes. Indeed, some argue that "*there is only one force that can effect a change… editors of major journals*" (Sedlmeier & Gigerenzer, 1989, p. 315). To submit to a particular journal, academics must conform to their guidelines with regards to scope, format, and any other requirements. Since journal editors or editorial boards have the power to create new guidelines which include specific comments on the use of statistics, they play a "*critical role… in the promotion of reforms in scientific practices*" (Giofrè et al., 2017, p. 10).

The first well-known example of editorial reform is the changes made at Memory & Cognition by Loftus during his time as editor (1994 – 1997). He discouraged the use of *p*-values and encouraged authors to rely on graphical presentations of data, with a strong emphasis on error bars (Loftus, 1993). While his policies had some success, with confidence interval and error bar rates increasing, and NHST-only papers decreasing, 1/3 of papers published during his time as editor still exclusively relied on NHST (Finch et al., 2004). In addition, very few authors discussed their error bars or confidence intervals in text, and once Loftus finished his term as editor, the occurrence of NHST-only papers increased once more. This suggests that his policies had a limited impact on both short-term and long-term behavioural changes. More recently, Psychological Science also modified their guidelines to encourage the use of the New Statistics, which was found to have a positive impact on reporting behaviour (Giofrè et al., 2017). However, the majority of authors still relied on NHST to form conclusions about studies, with effect sizes and confidence intervals interpreted in fewer than 20% of articles published under the updated reporting guidelines. These are two examples of the limited influence that open guidelines have on researcher behaviour, as neither of these cases required authors to comply with reporting standards.

Looking more specifically into statistical *requirements*, in 2005 the Journal of Clinical and Consulting Psychology became one of the first to require effect size and confidence interval reporting. One longitudinal review found that it had a marked impact on effect size reporting, rising from 65% of articles in 2004 to more than 90% of articles in 2008 (Odgaard & Fowler, 2010). While confidence interval use also increased, the change was much smaller, rising from 4% of articles in 2004 to just 38% in 2008 despite also being required by the journal. This suggests that even when statistics are required, enforcement from reviewers or editors is lacking, as effect sizes should typically always be accompanied by some measure of precision (e.g. confidence intervals), and so reporting rates should be similar if not identical.

It is not only individual journals that have taken steps towards statistical reform. As discussed in Chapter 1, the particularly prominent American Psychological Association (APA) has incorporated statistical advice into both their Publication Manual and their specially devised Journal Article Reporting Standards (JARS). As many journals adopt the APA standards for submissions, the actions of the APA are considered "*hugely influential*" on editorial practices (Fidler, 2002, p. 750). With particular reference to NHST, they ask that exact *p*-values are presented, along with effect size estimates and accompanying confidence intervals for each inferential test. The reporting standards also briefly mention statistical power, asking researchers to "*describe the sample size, power, and precision*" of their work, including the reporting of any power analyses used (Appelbaum et al., 2018, p. 7). The presence of these particular statistical concepts within the APA guidelines is one particular justification for focusing on these concepts both within Chapter 2 and across this thesis as a whole. This justification is similarly reflected in the contents of other popular organisational guidelines, such as those produced by the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) Network. While they were first developed to improve health research these standards have now been widely adopted by other disciplines where relevant (Simera et al., 2010). The various EQUATOR checklists typically instruct authors to report effect sizes and confidence intervals (or other estimates of precision) and to justify sample sizes, and also offer broader advice about not relying exclusively on *p*-values and avoiding selective reporting.

2.2.3 The Current Review

At present, typical examinations of journal guidelines have been conducted on a very small scale, evaluating just one or a few journals. It is not currently known how many journals have

either adopted organisational-level guidelines related to statistics, or created their own explicit policies related to statistics. This review seeks to identify the statistical guidelines across the top psychology journals, to assess the degree of statistical reform within the publishing industry. As researchers typically aim for the best possible journals for their work, the policies at these journals should have a significant impact on the statistical behaviour of researchers who wish to publish there. For this review, top psychology journals were identified using the Clarivate Journal Citation Reports™ published in June 2021, which assigns impact factors to each journal and produces a ranked database each year.

Review question: what are the statistical requirements or guidelines for submissions at the top 100 psychology journals (as ranked by impact factor), particularly with regards to NHST, effect sizes, confidence intervals, and statistical power?

2.3 Methodology

2.3.1 Ethics

An ethics checklist for this review was submitted and accepted in October 2020, in line with the ethical requirements at the University of Stirling for low-risk research projects.

2.3.2 Procedure

The Clarivate Journal Citation Reports™ published in June 2021 were used to create a database of journals categorised as 'Psychology' or any 'Psychology' sub-discipline, ranked by Clarivate impact factor. It should be noted that traditional impact factors are often criticised as not accurately measuring research quality (Krell, 2010). However, the impact factor of published work often forms part of academic employment evaluations, and so it is reasonable to presume that the majority of academic researchers will aim to publish in high-ranking journals in order to enhance their career evaluation and prospects. The Clarivate impact factors are typically advertised by journals to encourage authors to choose them for submission, which is why they have been used within this review.

The initial file generated by Clarivate consisted of 400 journals, which were then manually screened to produce a final list of the top 100 psychology journals, following the process in Table 2.1. Journals were excluded if they categorised themselves as an alternative discipline

on their website, such as management, medicine, or sociology. Multidisciplinary journals were kept in the database if they included psychology in the list of fields that they were open to publishing. Note that initially, only journals ranked from 1-200 were screened and refined to reduce the time required for this process; if this had resulted in fewer than 100 suitable journals, the remaining 200 journals would have also been screened.

Table 2.1

Refinement Process From Original Clarivate Database to top 100 Database

Process	Removed Journal Count	Remaining Journal Count
Original database downloaded		400
Screening of top 200 journals:	**200**	**200**
Non-psychology journals removed	49	151
Removal of qualitative-only journals	1	150
Removal of tutorial-only journals	1	149
Removal of review-only journals	28	121
Removal of journals ranked 101-121	21	**100**

Once the final list of 100 top journals was devised, author guidelines were sourced on each relevant journal website and examined for any mention of statistical requirements or recommendations. The primary investigation was to identify requirements related to the use of the New Statistics (specifically effect sizes and/or confidence intervals), mentions of statistical power and/or power analysis, and any comments or criticism related to *p*-values. Broader guidelines about justifying sample size and other statistical or methodological comments were also noted. An example of the coding strategy used to identify any guidelines related to effect size is presented in Table 2.2 (note that the same coding strategy was used for confidence intervals). A full table of codes can be found in Appendix A.

Table 2.2

Example Coding Strategy Used to Examine Author Guidelines for Effect Size

Code	Explanation
No	Statistical concept not mentioned in author guidelines, nor in any linked organisational guidelines
Rec	Statistical concept **recommended** explicitly by journal
Rec – G	Journal **recommends** adhering to guidelines which include statistic
Req	Statistical concept **required** explicitly by journal
Req - G	Journal **requires** adherence to guidelines which include statistic
Mixed	Requirements or recommendations differ by article type

Beyond statistical concepts, journals were also coded for guideline style (explicit/organisational), the presence of statistical resources (yes/no), the type of external guidelines shared (e.g. JARS, EQUATOR or other), and the accessibility of external guidelines (link provided/no link). This data was collected to illustrate the variety of guidelines that exist, and evaluate how well authors are supported in terms of accessibility and educational resources.

Once each journal had been fully examined, it was then categorised into one of four groups: no guidelines, mixed guidelines, statistical recommendations, and statistical requirements. Journals only had to present guidelines related to one statistical concept out of effect sizes, confidence intervals, NHST, or statistical power to be counted as having recommendations or requirements of some kind.

No Guidelines

As per the name, the category 'no guidelines' describes any journal which does not contain any guidelines related to statistical use or reporting, including no references to organisational guidelines, at the time this review was completed in October 2021.

Mixed Guidelines

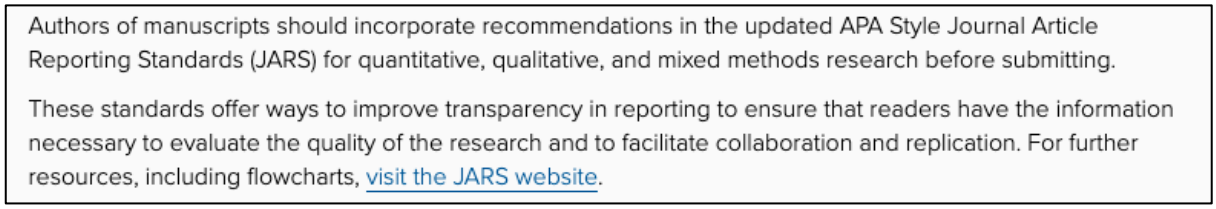
The category 'mixed guidelines' describes any journal which is mixed in one of two ways: (1) has a combination of requirements and recommendations for different article types or (2) only has statistical guidelines (whether they are requirements or recommendations) for particular article types (e.g. randomised controlled trials).

Statistical Recommendations

The journals categorised under 'recommendations' were those that either used vague language in their guidelines, such as '*please review the APA-JARS reporting guidelines*', or presented their instructions as '*authors should do [x]*', with no evidence that the reporting behaviour would be enforced. Recommendations could either be explicit, meaning that they were written within the guidelines for the journal, or organisational, meaning that authors were advised to follow some kind of common guidelines such as JARS. An example of organisational recommendations is shown in Figure 2.1.

Figure 2.1

Example of Organisational Recommendations From Psychotherapy (ranked #23)


Authors of manuscripts should incorporate recommendations in the updated APA Style Journal Article Reporting Standards (JARS) for quantitative, qualitative, and mixed methods research before submitting.

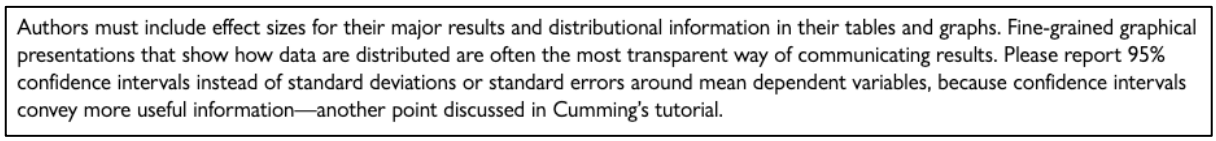
These standards offer ways to improve transparency in reporting to ensure that readers have the information necessary to evaluate the quality of the research and to facilitate collaboration and replication. For further resources, including flowcharts, visit the JARS website.

Statistical Requirements

Journals categorised under 'requirements' were any with statistical guidelines that were described with language such as '*authors must*' or '*authors are required to*', or where compliance with statistical guidelines had to be affirmed at the point of submission. In the same manner as recommendations, requirements could be explicit or organisational. An example of explicit requirements is shown in Figure 2.2.

Figure 2.2

Example of Explicit Requirements from Psychological Science (ranked #14)


Authors must include effect sizes for their major results and distributional information in their tables and graphs. Fine-grained graphical presentations that show how data are distributed are often the most transparent way of communicating results. Please report 95% confidence intervals instead of standard deviations or standard errors around mean dependent variables, because confidence intervals convey more useful information—another point discussed in Cumming's tutorial.

2.4 Results

This review found that journal guidelines related to the use of NHST, effect sizes, confidence intervals and statistical power varied widely, even within the same publishing houses. The full database associated with this chapter, which includes details such as publishing houses and impact factors, can be found at the OSF page associated with this thesis (found here). Impact factors of the 100 journals included in this review ranged from 20.652 to 3.603.

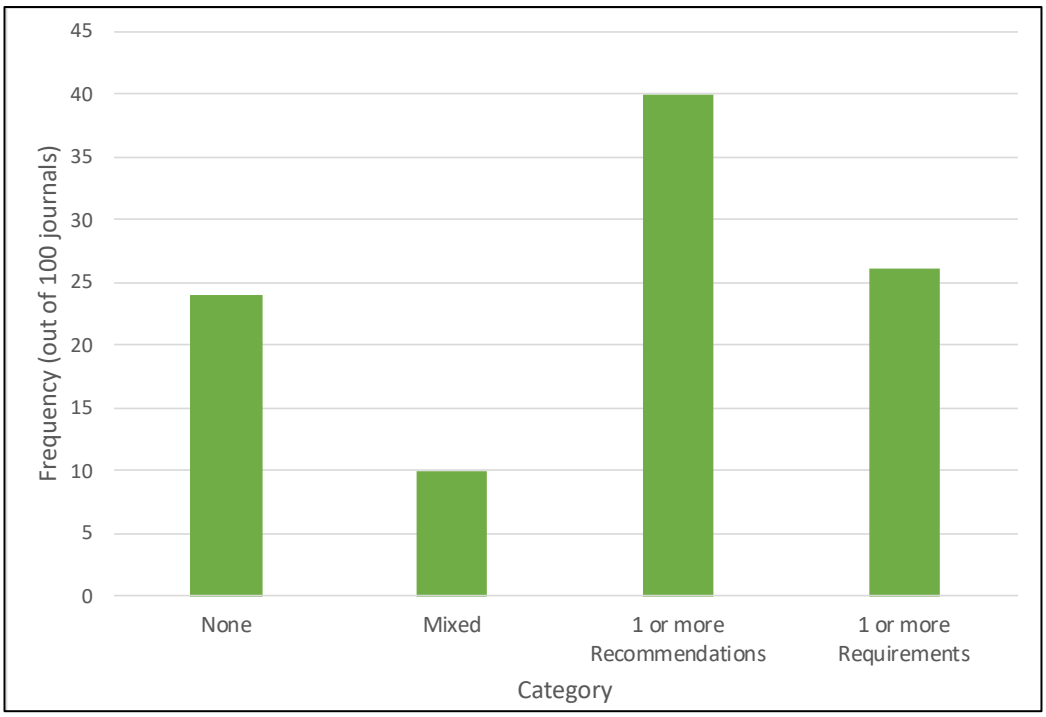
2.4.1 Overall Statistical Guidelines

Only 24 journals had no statistical guidelines for authors whatsoever, illustrated in Figure 2.3. Looking into this category further, a lack of guidelines was not associated with any particular impact factors. The highest-ranked journal with no guidelines was positioned at number 1 in the list of 100, and the lowest-ranked journal in this category was positioned at number 99.

The distribution of rankings and impact factors for each of the four categories is presented in more detail in Appendix A.

Figure 2.3

Summary of the Statistical Guidelines of the top 100 Psychology Journals



Of the 10 journals with mixed guidelines (guidelines which differed depending on article type), two only had recommendations for the presentation of clinical trials, and the rest (n = *8*) had a selection of requirements or recommendations across different articles. Forty of the top 100 journals had some form of statistical recommendations regarding one or more of NHST, power, effect sizes and confidence intervals, and the remaining 26 had at least one statistical requirement for submitted manuscripts (e.g. inclusion of effect sizes, or discussion of statistical power). Figure 2.4 provides more detail about the different statistical requirements and recommendations across the journals studied for this review.

2.4.2 The Contents of Statistical Guidelines

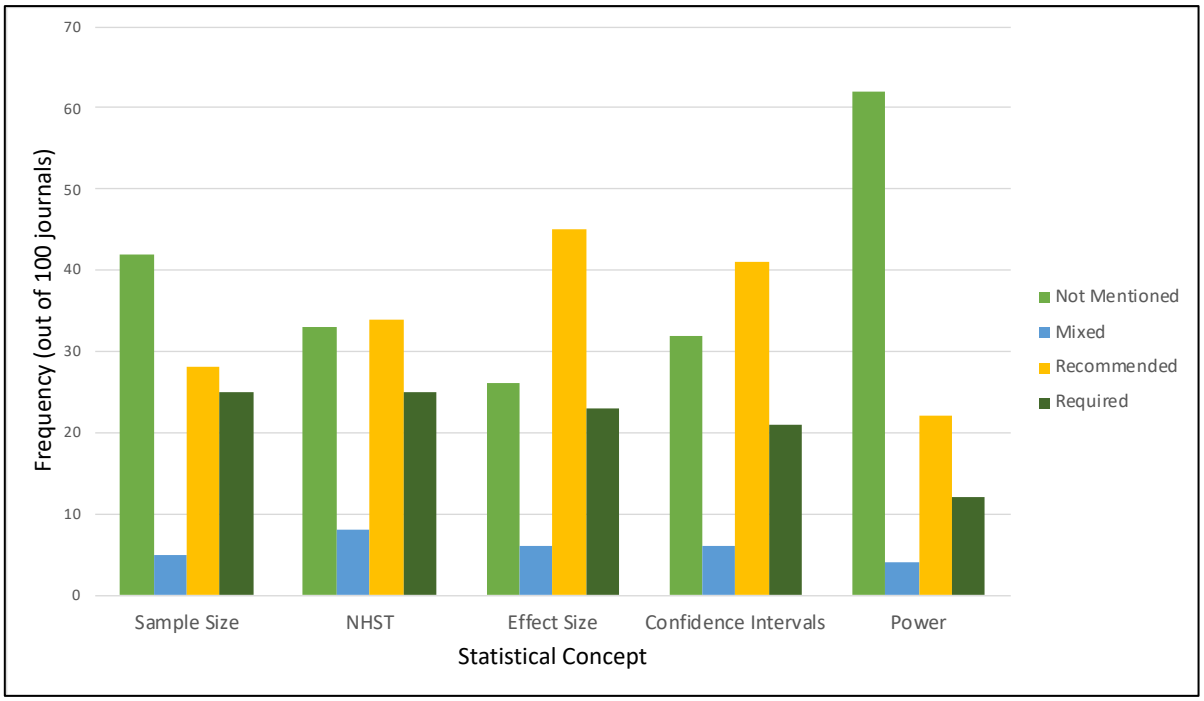
P-Values and NHST

Figure 2.4 shows that approximately two thirds of journals made some reference to *p*-values. Most commonly (in 53/67 cases), these were indirect and broad pieces of advice shared within organisational guidelines, such as the International Committee of Medical Journal

Editors' (ICMJE) advice to "*avoid relying on p-values*" (ICMJE, 2021, p. 17). A further 11/67 journals explicitly recommended (*n = 1*) or required (*n = 10*) that authors only reported exact *p*-values. Just 3/67 journals provided more detailed guidelines related to the use of NHST and *p*-values. One of these three journals explicitly required authors to correct for multiple testing, and two others, both published by Springer and associated with the Psychonomic Society, provided a detailed set of recommended best practices for working with NHST.

Figure 2.4

Frequency of Statistical Recommendations and Requirements



Note. The 'mixed' category indicates that only certain article types require or recommend a particular statistic.

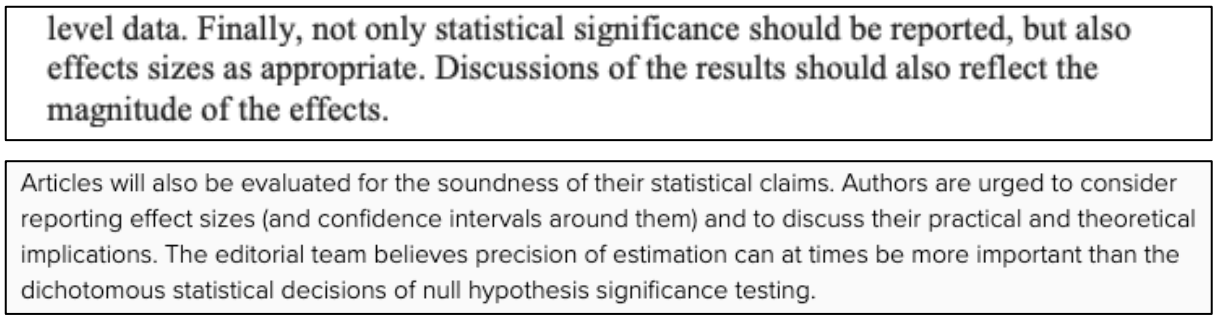
Effect Sizes

Effect sizes were the most popular statistic mentioned within author guidelines, neglected entirely by just 26/100 journals (as illustrated in Figure 2.4). Every journal which had an explicit guideline related to NHST (e.g. using exact *p*-values, or other more detailed recommendations) also explicitly encouraged or required effect sizes to be reported (which can be seen in the table in Appendix A). Where effect size reporting was recommended (*n = 45*), this was more likely to be through organisational guidelines (26/45) compared to through explicit author guidelines directly from the journal (19/45). In contrast, when effect size reporting was required (*n = 23*), it was more likely to be explicitly included in the journal

author guidelines (15/23) versus requiring the use of organisational guidelines (8/23). It should also be noted that just three out of 100 journals explicitly instructed authors to interpret or discuss their effect sizes, and that these requests were typically very short, as shown in Figure 2.5.

Figure 2.5

Requests to Discuss Effect Sizes, From Child Development (top) and the Journal of Experimental Psychology: General (bottom)



Confidence Intervals

As shown in Figure 2.4, slightly fewer journals encouraged or required confidence interval reporting compared to effect sizes, with one third of journals (32/100) neglecting them entirely. Twelve of the 21 journals *requiring* the use of confidence intervals included explicit directions in their author guidelines, while the remaining 8 required authors to adhere to organisational guidelines which include confidence intervals (which often contain the caveat that other intervals which represent uncertainty could also be used). Similarly to effect size use, when confidence interval use was recommended ($n = 41$), this was primarily through organisational guidelines (32/41), compared to explicit recommendations within individual journal guidelines (9/41). In the remaining six journals with mixed guidelines, all six directed authors to organisational guidelines rather than providing any explicit individual instructions. None of the 100 journals examined within this review appeared to encourage authors to discuss their confidence intervals when evaluating their findings.

Power and Sample Size

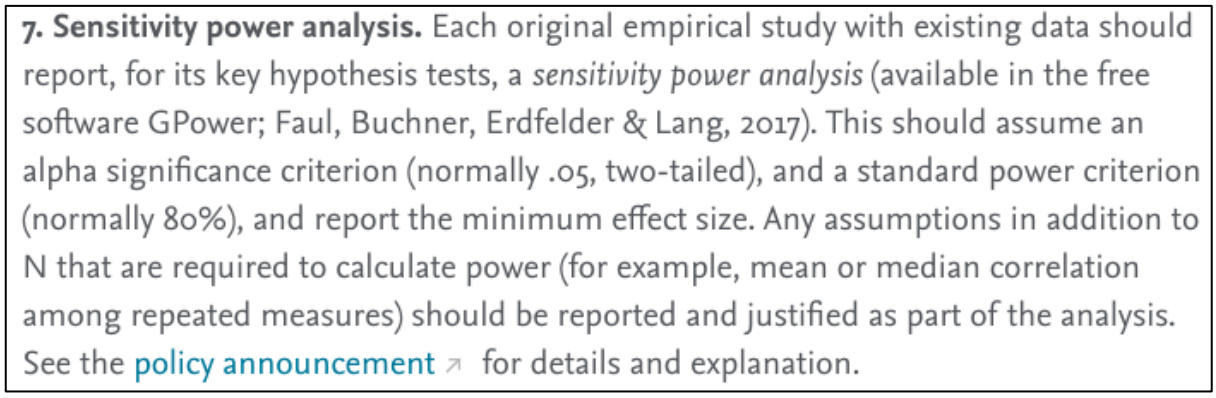
More than half ($n = 58$) of the journals in this review asked for some form of sample size justification from authors, although often this was not explicitly connected to statistical power. Twenty-five journals required sample size justifications for all quantitative research, with an additional four of the five 'mixed' journals requiring it for randomised controlled

trials or registered reports. A further 28 journals encouraged authors to explain or evaluate their sample size for all studies; while the fifth mixed journal encouraged it for some studies.

Just 38 of these 58 journals mentioned statistical power with regards to sample size. Overall, just one journal (*Personality and Social Psychology Bulletin*) required power analyses for all quantitative submissions; while a further two journals specifically required sensitivity power analyses instead of, or in addition to, a priori power analyses (example shown in Figure 2.6). More commonly, journals either explicitly required ($n = 2$) that authors discussed the power of their study when discussing their findings; or required that authors follow JARS ($n = 7$), which instructs authors to "*describe sample size and power"* (Appelbaum et al., 2018, p. 7). It should be noted that two of the journals which require power analyses of some kind (a priori or sensitivity) also explicitly banned the use of post hoc or 'observed' power analyses, which are widely criticised (e.g. Yuan & Maxwell, 2005; discussed further in Chapter 7).

Figure 2.6

Author Guidelines Related to Statistical Power From the Journal of Experimental Social Psychology


7. Sensitivity power analysis. Each original empirical study with existing data should report, for its key hypothesis tests, a *sensitivity power analysis* (available in the free software GPower; Faul, Buchner, Erdfelder & Lang, 2017). This should assume an alpha significance criterion (normally .05, two-tailed), and a standard power criterion (normally 80%), and report the minimum effect size. Any assumptions in addition to N that are required to calculate power (for example, mean or median correlation among repeated measures) should be reported and justified as part of the analysis. See the policy announcement ↗ for details and explanation.

Note. For this journal, sensitivity power analyses had to be confirmed upon submission.

Statistical power was more frequently incorporated into recommendations, with 15 journals strongly encouraging the use of an a priori power analysis, recommending power discussions ($n = 1$) or encouraging authors to follow JARS ($n = 6$). Within the five mixed journals which referenced sample size, one recommended following JARS for some study types, and three required the use of power analyses for registered reports or randomised controlled trials (just one of the five 'mixed' journals asking for a sample size justification did not have a subsequent guideline related to statistical power).

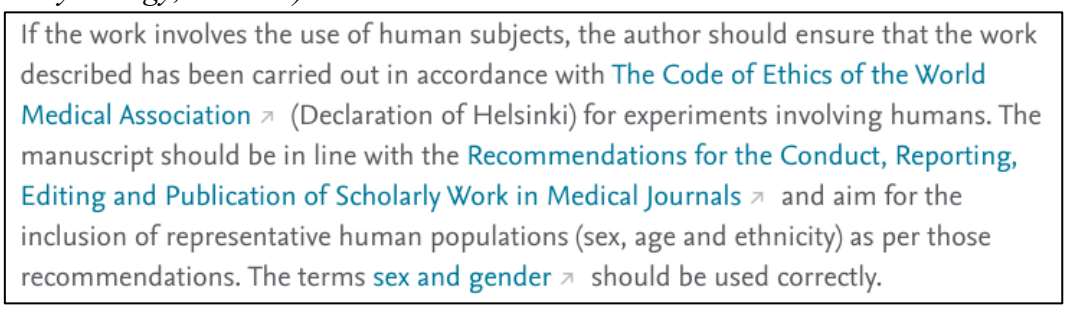
Organisational Guidelines

The most popular organisational guidelines that have been adopted by the top 100 psychology journals are those that are within the EQUATOR network, which was originally conceived to improve health research. Twenty-seven of the 100 journals studied here required or encouraged authors to follow one or more of the EQUATOR checklists, with the most popular being CONSORT for randomised controlled trials.

The APA's JARS guidelines also appeared frequently, mentioned in the author guidelines from 20 of the 100 journals. Use of these guidelines does not appear to be limited to just the APA publishing house journals, with four journals from Wiley, Taylor and Francis, and COP Madrid also incorporating them into their instructions to authors. Another equally popular option was the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals, published by The International Committee of Medical Journal Editors (commonly abbreviated to the ICMJE guidelines). They advise authors to "*quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)*" and to "*avoid relying solely on statistical hypothesis testing, such as p-values, which fail to convey important information about effect size and precision of estimates*" (ICMJE, 2021, p.17). Twenty journals requested that authors produce a manuscript in line with these guidelines, although this was almost always presented using vague language as shown in Figure 2.7.

Figure 2.7

Example of how the ICMJE Guidelines are Presented by Journals (shown in Journal of School Psychology, Elsevier)


If the work involves the use of human subjects, the author should ensure that the work described has been carried out in accordance with The Code of Ethics of the World Medical Association ↗ (Declaration of Helsinki) for experiments involving humans. The manuscript should be in line with the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals ↗ and aim for the inclusion of representative human populations (sex, age and ethnicity) as per those recommendations. The terms sex and gender ↗ should be used correctly.

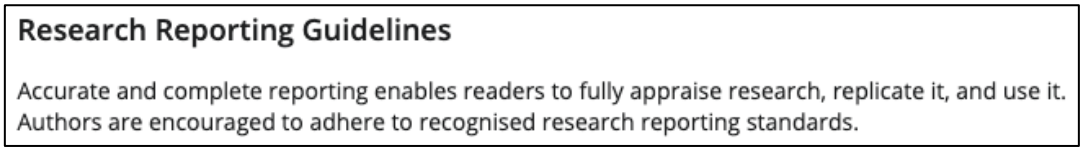
2.4.3 Accessibility and Support

While most journals provided URLs when requesting that authors follow organisational guidelines, five failed to do so, leaving authors to search online for themselves. Of these five,

two requested that authors follow the APA's JARS without providing links, and three Wiley journals vaguely encouraged authors to adhere to reporting standards, without further detail (as illustrated in Figure 2.8).

Figure 2.8.

Example of Limited Information Included in Wiley Author Guidelines

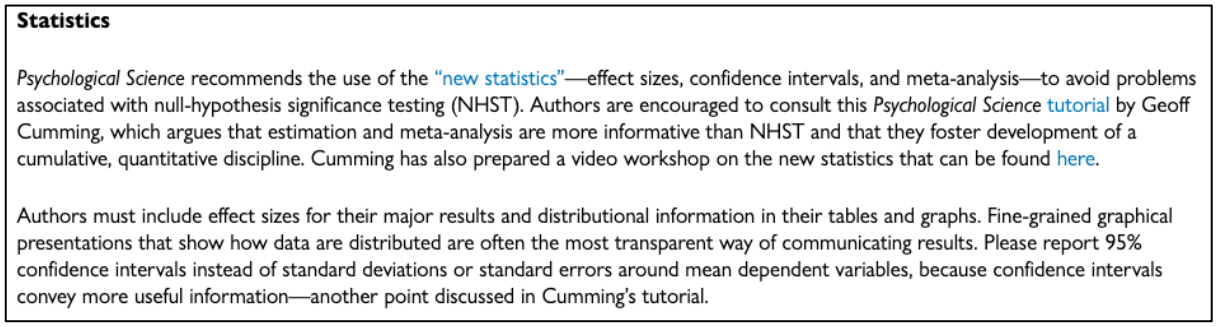


Research Reporting Guidelines

Accurate and complete reporting enables readers to fully appraise research, replicate it, and use it. Authors are encouraged to adhere to recognised research reporting standards.

Figure 2.9

Psychological Science Statistics Support Within Author Guidelines



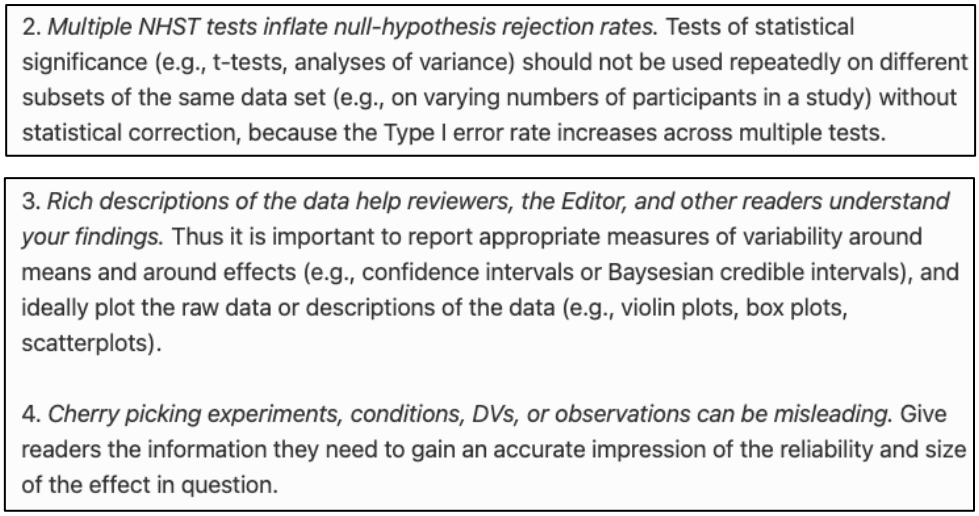
Statistics

Psychological Science recommends the use of the "new statistics"—effect sizes, confidence intervals, and meta-analysis—to avoid problems associated with null-hypothesis significance testing (NHST). Authors are encouraged to consult this *Psychological Science* tutorial by Geoff Cumming, which argues that estimation and meta-analysis are more informative than NHST and that they foster development of a cumulative, quantitative discipline. Cumming has also prepared a video workshop on the new statistics that can be found here.

Authors must include effect sizes for their major results and distributional information in their tables and graphs. Fine-grained graphical presentations that show how data are distributed are often the most transparent way of communicating results. Please report 95% confidence intervals instead of standard deviations or standard errors around mean dependent variables, because confidence intervals convey more useful information—another point discussed in Cumming's tutorial.

Figure 2.10

Two Examples From the Behavior Research Methods (Springer) Statistical Guidelines



2. *Multiple NHST tests inflate null-hypothesis rejection rates.* Tests of statistical significance (e.g., t-tests, analyses of variance) should not be used repeatedly on different subsets of the same data set (e.g., on varying numbers of participants in a study) without statistical correction, because the Type I error rate increases across multiple tests.

3. *Rich descriptions of the data help reviewers, the Editor, and other readers understand your findings.* Thus it is important to report appropriate measures of variability around means and around effects (e.g., confidence intervals or Baysesian credible intervals), and ideally plot the raw data or descriptions of the data (e.g., violin plots, box plots, scatterplots).

4. *Cherry picking experiments, conditions, DVs, or observations can be misleading.* Give readers the information they need to gain an accurate impression of the reliability and size of the effect in question.

With regards to statistical support, just eight journals provided some form of statistical resources for authors. Typical resources took the form of links to published journal articles about statistics, suggested values to use in an a priori power analysis, or links to tutorials, such as the guidance offered by Psychological Science which is shown in Figure 2.9. Two of

these eight journals belonged to the Springer group, and both provided detailed information and guidance related to using NHST in the form of a detailed 6-item list (shown in Figure 2.10).

2.5 Discussion

This review contributes to the literature by providing one of the first systematic insights into the statistical guidelines of psychology journals, highlighting the prevalence of recommendations and requirements that relate to statistical reform in psychology. Just one quarter of the 100 journals examined had no statistical guidelines whatsoever. It is clear that large organisations have had a noticeable impact on editorial policies, as many journals have adopted guidelines such as the APA's JARS, or the EQUATOR checklists. As journal guidelines are described as playing a "*critical role… in the promotion of reforms in scientific practices*" (Giofrè et al., 2017, p.10), the growing adoption of policies that encourage researchers to report more than just *p*-values is likely to positively impact the transparency of published research.

2.5.1 The Variety of Statistical Guidelines

This review found that, while many journals have at least one comment in their author guidelines that relates to statistics in some way, the content and format of statistical guidelines varies enormously. Journals ranged from providing single sentences about one concept through to providing clearly signposted "Statistical Guidelines" sections with detailed instructions for authors. Despite the use (or misuse) of NHST and *p*-values being at the heart of statistical reform, few journals made explicit comments about how to use NHST appropriately, instead encouraging authors to support their *p*-values with effect sizes and confidence intervals, or to discuss the statistical power of their research.

With regards to the 'New Statistics', effect sizes appear to be the most popular addition to author guidelines. This reflects the trends seen across the literature, where effect sizes are far more widely reported than confidence intervals (e.g. Fritz et al., 2013). However, just three of the reviewed journals explicitly asked that authors *discuss* their effect sizes (and when doing so, provided little detail beyond considering practical importance), with no apparent instructions to evaluate confidence intervals anywhere. Previous reviews have demonstrated

that, even when reported, effect sizes and confidence intervals are interpreted less than 20% of the time (Giofrè et al., 2017). Taken together, these findings demonstrate that authors need more encouragement to make use of all statistics when interpreting their data.

Twenty-one out of 100 journals either encouraged or required power analyses to be used and reported, including two references to using sensitivity power analyses either instead of or in addition to a priori calculations. The emergence of sensitivity power analyses is promising, because they are often advocated for as a sensible way to evaluate the power of a study by indicating the smallest effect sizes that could reliably be detected with a particular sample size (Bacchetti, 2010), and do not require estimations of effect sizes. Indeed, this could be a valuable recommendation for more editors and reviewers to adopt, particularly to discourage the use of the uninformative post hoc power analyses (as discussed in Chapter 1, Section 1.4).

2.5.2 The Limitations of Statistical Guidelines

While it is certainly positive that so many journals are beginning to adopt statistical guidelines, more attention should also be paid to the accessibility of these instructions. In many cases statistical guidelines were not noticeably signposted with subheadings, while organisational guidelines were often only mentioned briefly, and so could be easily missed by authors. In addition, frequently authors were directed to generic webpages for organisational guidelines, which creates a burden on authors to search through multiple webpages to find the documentation they need to use. The language used in many of these guidelines also means that authors are effectively free to adopt or ignore statistical recommendations as they wish. For example, both the ICMJE and JARS guidelines clearly position themselves as recommendations, leaving decisions down to author discretion. Even when a journal indicates that authors ‘must comply with JARS’, it is not always clear whether they mean simply take advice from JARS, or actually treat JARS as a comprehensive reporting checklist to be followed.

This flexibility could be perceived as both a good and a bad thing. It does not encourage those who are not personally interested in statistical quality to adopt changed behaviours; but it also stops short of policing the behaviour of researchers. However, after decades of encouragement from various sources, such as the advice of Cohen (1990) or the various reports and guidelines shared by the APA (Wilkinson, 1999; Appelbaum et al., 2018), reviews indicate that reporting behaviours are not yet transformed (e.g. Fritz et al., 2013).

While more contemporary reviews are needed to look at recent change, the existing evidence suggests that plenty of researchers might need the 'stick' aspect of a carrot-and-stick approach to provide extrinsic motivation for statistical reform. This is further reflected in the lack of *interpretation* of estimation-based statistics, both in the guidelines provided by journals but also identified through reviews of the literature (e.g. Giofrè et al., 2017), which implies that many researchers have not yet found value in the use of estimation statistics to critically evaluate findings (a potential 'carrot' incentive).

It is also important to acknowledge that changing reporting practices is only the first step in true statistical reform, as writing numbers on a page does not necessarily mean that the numbers themselves are understood, or indeed have been produced without errors. Simply asking researchers to follow particular instructions does not mean that they have sufficient knowledge to do so correctly. Plenty of research exists which demonstrates that *p*-values are both misunderstood conceptually, and are also often presented with incorrect test statistics and mathematical details (e.g. Lyu et al., 2018; Nuijten et al., 2016). Now, research needs to establish that the same mistakes are not being made with regards to effect sizes, confidence intervals, and statistical power.

Of the 76 journals reviewed here that do have some kind of statistical guidelines, only a few offered some form of educational resources for authors. It would be unreasonable to place the burden of education on journals: this is not their role in the research ecosystem. However, journals do have the advantage of already providing submission guidelines for researchers to follow; and so it would not be an unreasonable jump to consider using these widely-accessed spaces as a platform for resources and tutorials, too. For instance, within the author guidelines of the Psychonomic Society journals (Figure 2.10, section 2.3.3) are a series of educational instructions related to statistical analysis, combining advice with statistical facts. This provides context to their instructions, so that their guidelines are an opportunity to learn and make educated choices when working with statistics. More minimally, Psychological Science presents links to tutorials and videos for authors, which requires very little effort on the part of the journal, but provides easily accessible help for authors; an approach which could be adopted by other journals. However, it is important to note that despite the enthusiasm of Sedlmeier and Gigerenzer ("*there is only one force that can effect a change… editors of major journals*"; 1989, p. 315), individual researchers are also responsible for changing their own practices and improving their statistical evaluations, regardless of the

instructions given by any singular journal. As a growing number of psychology courses are teaching students about the limitations of NHST, and concepts that can support or replace it (TARG Meta-Research Group, 2020), the next generation of researchers should be better equipped to independently make improved statistical decisions.

2.5.3 Review Limitations

There are two notable limitations to this work. Firstly, it is important to bear in mind that author guidelines will change over time as organisational policies and editorial staff come and go. The number of top journals with statistical guidelines will look different every year, although if one is optimistic, then perhaps the number of statistical guidelines will increase over time as psychology moves forwards. Secondly, this review has only sampled the top 100 journals ranked by a traditional impact factor, meaning there are numerous other psychology journals publishing research that have not been studied, which could have any range of guidelines and educational resources for authors.

2.6 Conclusion

It can be seen here that in many cases, authors publishing in the top 100 psychology journals will have to widen their statistical reporting beyond *p*-values, and take a more critical approach to sample size planning, to comply with author guidelines. Encouraging reporting is a positive step towards improving the quality of research dissemination, but researchers must also *understand* the statistics they are using, to avoid the mistakes that have been identified with the historic use of *p*-values. The lack of support resources, such as educational tutorials, indicates that to comply successfully with most journal guidelines, the 'ideal' researcher will need their own reasonable understanding of each of these statistical concepts. Research into the use and understanding of the three key statistical concepts of this thesis (effect sizes, confidence intervals, and power) is presented in Chapters 3-7.

Part 1: The New Statistics

Chapter 3: The Use and Knowledge of Effect Sizes in Psychology

3.1 Abstract

Background: An effect size is a value which quantifies the size of a phenomenon of interest. Whilst historically they have not been widely used, recent crises and reform efforts have renewed calls for their inclusion in psychological research. Reviews of the literature show that use is increasing, but very little is known about individual experiences or barriers to use. Similarly, little is known about how well researchers understand effect sizes. This chapter provides new insights into effect size use and knowledge.

Methods: An online questionnaire was used to examine effect size use, barriers to use, and effect size knowledge in a sample of 247 psychology researchers. Participants were asked about their experience of using, or not using effect sizes, including software preferences, publishing experience and explanations for not using effect sizes. As a test of knowledge, participants were asked to answer five novel true-false statements and also to define the term 'effect size' in their own words.

Results: Self-reported effect size use was high, with only 9.4% of participants in this sample never using effect sizes. One third of participants scored full marks on the true-false knowledge test, with broadly high scores across the full sample, but many misconceptions were apparent when participants were asked to define the term effect size. Barriers to using effect sizes include lack of motivation and lack of knowledge.

Conclusions: While external incentives such as regulations are likely to improve statistical behaviour, researchers must also be shown that effect sizes are a useful tool beyond a check-box exercise, in order to increase individual motivation to change statistical behaviour. Furthermore, more education is needed to demonstrate that there are a wide variety of standardised and unstandardised effect sizes that suit many different research questions, which could enhance transparent and detailed reporting across all of psychology.

3.2 Introduction

An effect size is defined as any value which quantifies the size of a phenomenon of interest. As discussed in Chapter 1 of this thesis, effect sizes are a key statistic within the estimation approach, as they are used to answer the question of 'how big is the effect?'. The review presented in Chapter 2 highlighted that effect sizes are the most frequently requested statistic within journal author guidelines, and are often now required for quantitative research. Given their popularity, they are the first topic that this thesis will focus on. This chapter presents the findings of a questionnaire study, which examines the use and knowledge of effect sizes in a sample of psychology researchers. Note that in this thesis, the term 'effect size' is used in the broadest sense to represent any kind of effect size index, including both standardised values (e.g. Cohen's *d)* and unstandardised values (e.g. a mean difference).

3.2.1 Effect Sizes in Psychology

Historically, psychological research has lacked effect size reporting beyond those automatically calculated as part of a statistical test, such as Pearson's *r* (generated when testing correlations). This lack of additional statistical reporting can be attributed, in part, to the prevalence of null hypothesis significance testing (NHST), which has historically been prioritised as the key statistical criteria used to judge and publish research articles. NHST, and the *p*-values that it produces, are used to evaluate data with a dichotomous significant or not-significant approach. This often discourages thoughtful analysis, and provides very little information about the phenomena being studied in many cases. In contrast, effect sizes offer measures of magnitude, which are often useful evidence to answer the research questions that are being studied.

The use of effect sizes has grown noticeably over the past two decades, particularly with encouragement from prominent organisations. For example, at the end of the 20th century the American Psychological Association (APA) Task Force on Statistical Inference issued a report stating that academic psychologists should "*always present effect sizes for primary outcomes",* describing effect sizes as "*essential to good research*" (Wilkinson, 1999, p. 602). They further reinforced this by incorporating similar messages into their Publication Manuals (e.g. APA, 2001). One review suggests these efforts have been successful, as effect size

reporting increased to approximately 55% of published articles post-1999, compared to 30% in those published pre-1999 (Peng et al., 2013).

However, it is important to note that reporting effect sizes does not represent true statistical reform. One issue with NHST is that it has been widely misunderstood by researchers, resulting in incorrect interpretations (e.g. Haller & Krauss, 2002) and erroneous reporting (e.g. Nuijten et al., 2016). To be used effectively, effect sizes must not only be reported, but also understood and interpreted; which means that researchers must be equipped with sufficient knowledge to do so correctly.

3.2.2 Researchers and Effect Sizes

At present, very little is known about effect size knowledge and understanding across the psychology researcher population. Despite many reviews of effect size reporting in the published literature (e.g. Fritz et al., 2013; Peng et al., 2013), only a small amount of research has directly surveyed psychologists about their use and knowledge of effect sizes. One survey of 472 Spanish academic psychologists found that 87.1% claimed to know about effect sizes, although only 68.4% could name an effect size statistic (Badenes-Ribera et al., 2016). While 40.7% of participants reported using effect sizes 'quite often', nearly a quarter of the sample reported using effect sizes never, or very infrequently. A subsequent replication, which surveyed 159 Italian psychologists, found even less familiarity with effect sizes: 81.8% of participants claimed to know about them, but only 44.7% could name an effect size statistic (Badenes-Ribera et al., 2018). Within this study, only 35.9% of participants reported using effect sizes 'quite often', with nearly one third of the sample using effect sizes 'never' or 'very little'. As yet, there is little evidence that psychologists have a firm understanding of effect sizes.

3.2.3 Chapter 3 Overview

The intention of this study was to contribute new data on the use and knowledge of effect sizes in the psychology researcher population, through a combination of quantitative and qualitative data. The specific objectives were as follows:

Objective 1: To examine effect size use including software preferences and publishing experience, along with explanations for not using effect sizes, to capture individual perspectives and experiences.

Objective 2: To use a novel version of the true-false testing method to explore basic conceptual knowledge of effect sizes in this population, and how they are perceived in relation to other statistical concepts such as NHST and sample size.

Objective 3: To explore perceptions of what the term 'effect size' means to different psychology researchers, and identify any misconceptions of effect sizes, by asking participants to define the term in their own words.

3.3 Methodology

3.3.1 Ethics

This study received ethical approval from the General University Ethics Panel (GUEP #829 19-20) and adhered to the Code of Human Research Ethics guidelines of the British Psychological Society (BPS) (BPS, 2014). Documentation can be found in Appendix B.

3.3.2 Sampling and Inclusion Criteria

Due to the exploratory nature of this study, a power analysis was not deemed suitable for identifying a recommended sample size. Participants were recruited using opportunity sampling to capture as many participants as possible during a month-long sampling window (10th February to 10th March 2020). The study was advertised on Twitter and Reddit, and shared on academic mailing lists. Within the University of Stirling, the study URL was distributed to the psychology staff and psychology PhD student mailing lists, and externally the study URL was distributed to multiple JISCMail lists. An example advertisement is shared in Appendix B.

Any self-identifying psychology researcher actively involved in quantitative research in any location was eligible to take part in this study, including PhD students and MSc-by-Research students. Psychologists outside of traditional academia, but involved in research to any extent (e.g. clinical psychologists and similar) were included in this population. Undergraduate psychology students were not eligible as they typically do not publish research.

3.3.3 Materials

Data for this study was collected via an online questionnaire. All questions were developed by the researcher, and are detailed below. The questionnaire was first piloted with a professor, a lecturer, and two PhD students, to identify any issues with question clarity and questionnaire flow from a variety of perspectives. Subsequently, response boxes for two questions were modified to ensure that they were big enough for longer feedback; no other changes were advised. The full questionnaire can be found in Appendix B.

Definition of Effect Size. All participants were asked to provide a definition of the term *effect size* in their own words, or alternatively write *I don't know* in the response box. This was asked as a free-text question to capture the different perceptions of an effect size that exist in the psychology researcher population, along with identifying mistakes and misconceptions that may exist.

Personal Use of Effect Sizes. Similarly to the Badenes-Ribera et al. (2016) study, participants were asked if they currently calculate effect sizes in their own research, with response options *yes, no* or *not always*. Participants who responded *not always* or *no* were given the opportunity to explain why not, if they wished to do so.

Participants who responded *yes* or *not always* were then asked what software they used for effect size calculation, to capture trends in technology use. They were also asked if they had included any effect sizes in any of their published papers or pre-prints. To examine whether journal regulations impact the adoption of effect sizes in published work, participants who responded *yes* were asked to reflect on their first published effect size and report whether *journal requirements*, *personal decisions*, or *a combination of both* influenced their decision to include effect sizes in a paper.

Importance of Effect Sizes. The perceived importance of effect sizes was measured with the question "*how important do you feel effect sizes are in psychological research?".* Participants answered using a Likert-style response item, with four options from *not important at all (1)* to *very important (4)*, with an alternative fifth option of *I don't know*. This question was used as a measure of current attitudes towards effect sizes.

Training in Effect Sizes. Participants were asked whether they felt that they had been provided with, or had access to, sufficient training about effect sizes, with the response options *yes, no* and *prefer not to say*. They were then asked if they would make use of training, if it were made available, with options *yes, maybe, no* or *prefer not to say*. Participants who opted for *maybe* or *no* were shown a follow up open-ended question asking them to explain their response, if they were willing to do so. Asking participants for their feedback was a way to capture potential barriers to engaging with training, which may improve the accessibility of future training.

Knowledge of Effect Sizes. A true-false series of statements was used to examine basic knowledge of effect sizes, looking for levels of understanding and possible misconceptions. The true-false approach was selected as there is minimal current research into effect size knowledge, and this approach has been used in other statistical contexts, particularly for *p*-values (e.g. Haller & Krauss, 2002) and confidence intervals (e.g. Hoekstra et al., 2014). In brief, this method presents participants with a series of statements about a concept, and asks them to rate each one as being true or false. Traditionally, all statements are false, and so labelling any as true is used as evidence of misunderstandings.

It should be noted that this method has been criticised for several reasons. The first issue is that restricting participants to only answering true or false means that true misinterpretation cannot be distinguished from mere guesswork; and the second is that only providing false statements fails to allow participants to indicate any correct knowledge that they might possess (Garcia-Pérez and Alcalá-Quintana, 2016). As recommended by Garcia-Pérez and Alcalá-Quintana (2016), the study reported in this chapter has used a modified version of the traditional true-false approach, by including a third 'I don't know' response option for each statement, to reduce guessing. The statements used in this study were also developed to incorporate one true statement (instead of all being false), to allow participants the opportunity to demonstrate awareness of a plausible definition of an effect size. In addition, an odd number of statements was presented so that participants were not potentially mislead into anticipating an equal number of true and false statements.

Typically, the statements used in studies like this are based on common misconceptions, or *fallacies*, about the topic of interest. However, the limited research into effect sizes and how well they are understood restricts this option. Statement development, led by the researcher,

was based on wider reading around the subject. The final set of five statements is found in Table 3.1, with the correct response (*true* or *false*) indicated. Statement 1 presents a correct definition of effect size, based on Kelley and Preacher, who define effect size as "*a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest*" (2012, p. 137). Statements 2 and 3 are inspired by reviews of effect size descriptions and interpretation (Funder and Ozer, 2019; Morris, 2020), while Statements 4 and 5 relate to the *magnitude fallacy,* which describes the false belief that statistically significant results indicate the presence of large effect sizes (Kline, 2004; Kühberger et al., 2015).

Table 3.1
Five True-False Statements Presented to Participants

Statement	True or False?
Effect sizes express the magnitude of the influence one variable has on another	True
A larger sample size will result in a larger (stronger) effect size	False
If you are doing high quality research, your aim is to find the largest effect sizes possible	False
A statistically significant *p*-value corresponds to a medium or large effect size	False
A small effect size indicates that the null hypothesis should fail to be rejected	False

Comment on Scale Validation

It is important to note that these are novel statements, which have been combined into an unvalidated scale. As this is exploratory research, validation was not deemed necessary in the context of this single study, although a 5x5 correlation matrix for this scale is shared in Appendix B. Scale development, statement choices and potential future use will be evaluated later in this chapter.

Demographics. Limited demographics were collected for this study. It was decided that age, gender and ethnicity were not relevant to the objectives of the research. Participants were asked to report their academic job position, and were given a list of UK-based academic job roles with descriptions of similar international job titles in brackets, such as Professor (equivalent to tenure-track top level positions), with an "Other" free-text box available to

report other roles. They were also asked to report the country their academic position was based in and the sub-field of psychology they would categorise themselves into (note that both of these questions were asked as free-text items). Participants were finally asked "*are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar?*", with response options *yes, no* or *prefer not to say*. This variable has been given the abbreviated name of 'Open Science' for brevity.

3.3.4 Procedure

Qualtrics (Qualtrics, Provo, UT) was used to deliver the questionnaire online. Participants were first shown an information sheet which described the inclusion criteria and provided a short overview of the study. They then gave consent via a digital consent form, which repeated the inclusion criteria to confirm that participants were suitable to take part. It was signposted in the information sheet that all questions were optional and could be skipped if preferred. Participants could exit the questionnaire at any point and had seven days to re-enter the questionnaire and finish it, if they wished to do so.

The questionnaire questions were shown in the order listed within the Materials section (3.3.3), in an order designed not to prime responses on later questions or deter participants from finishing the questionnaire. For example, participants were first asked to define effect size to capture their baseline knowledge about effect sizes before seeing more questions about them. The true-false statements were presented near the end, in case they deterred participants from carrying on with the questionnaire in order to avoid any further potential questions about statistics. Participants were asked to provide their email address at the end of the questionnaire if they were interested in hearing about follow-up research, which was explicitly described as optional. Email addresses were held separately from questionnaire responses to preserve anonymity.

3.3.5 Data Handling

Overall, 326 participants started the questionnaire, but 75 responses were removed because the participant did not move beyond the consent page. One further response was removed because the participant did not appear to give consent, which was attributed to a Qualtrics error in their 'Force Response' functionality. A further three participants were removed due

to only answering a random selection of questions and providing no complete responses to any section of the questionnaire. The resulting data set consisted of data from 247 participants, details of which are found in section 3.3.8.

3.3.6 Quantitative Analysis

The *personal use of effect sizes, software use, perceived importance, training,* and *true-false statement* data were all analysed using descriptive statistics, generating frequencies and averages where appropriate. Exploratory null hypothesis statistical testing was used to investigate potential demographic group differences. Note that for chi-squares, due to small expected frequencies, the 'job role' variable was compressed into four categories: pre-doctoral participants, doctoral students, post-doctoral researchers, and professors, and only the 'yes' and 'no' groups of the open science variable were included in these analyses. The five true-false scale items were also combined to create an exploratory score of effect size 'knowledge' for each participant. Quantitative analyses were computed using Jamovi (The Jamovi Project, 2021) or Microsoft Excel.

3.3.7 Qualitative Analysis

Qualitative data from three questions was analysed using basic content analysis, which is defined as a "*research technique for the objective, systematic and quantitative description of manifest content of communication"* (Berelson, 1952, p.18, as cited in Drisko and Maschi, 2015, p. 3). A basic content analysis is a method of summarising qualitative data using a combination of qualitative coding and quantitative counting to form categories (Drisko and Maschi, 2015), as illustrated in Figure 3.1.

Figure 3.1

Example of Basic Content Analysis, Using an Inductive Approach

Meaning Unit (Response)	Code(s)	Suggested Category	Final Category
I'm not totally sure of my statistical knowledge	Lack of knowledge	1	1
Mostly habit, but I am starting to calculate effect sizes more frequently	Habit	3	3

In the work presented in this chapter, and subsequently in this thesis, when content analysis is used, it is used at the manifest level. This means focusing on the literal words shared by

participants, rather than trying to uncover deep interpretations. Each individual response to a question represents a single unit of meaning, which was coded in the context of the particular question asked of participants.

Content analysis can take two approaches: inductive coding ('bottom-up'), which identifies ideas within the data, or deductive coding ('top-down'), which uses preconceived ideas along with a readthrough of the data to first create a list of codes, which are then applied to the data (and expanded if needed). A combination of both of these approaches are used here. Data related to not using effect sizes, or not being interested in training, were both analysed inductively. In contrast, definitions were analysed both deductively and inductively, as described below.

Content Analysis: Definitions of Effect Size

A content analysis was carried out twice to examine participant definitions of effect size. First, a deductive coding process was used to categorise each definition as *acceptable, incorrect, or shows some understanding* (as detailed below). The label *acceptable* was used, instead of *correct*, as the literature demonstrates that there are broad ways to describe an effect size.

Kelley and Preacher's (2012) definition of effect size, "*a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest*" (p. 137), generated in their review of the effect size literature, was used as a guide for this categorisation process; along with Watt and Collins' (2019) definition, "*a numerical description of the strength of the relationship [between variables]*" (p. 70). Definitions were rated as *acceptable* if they demonstrated a correct understanding of the concept of effect size using at least two of the three key elements in the above definitions: magnitude, measure and phenomenon (or suitable synonyms). The deductive content analysis process was used to identify the prevalence of each element. Categorising a definition as acceptable was also determined by what it did *not* contain, such as vague language, the mention of incorrect concepts (e.g. power), or defining an effect size under specific circumstances such as only being the product of an intervention, or only being a standardised measure.

Definitions of effect size that described other incorrect concepts (e.g. power, or degrees of freedom) were categorised as *incorrect,* and all definitions considered 'in between'

acceptable and *incorrect* were categorised as *shows some understanding*. All definitions categorised as *shows some understanding* were subsequently further divided into the subcategories *vague definitions* or *overly specific definitions.* Vague definitions provided little detail but typically acknowledged that an effect size was the size of an effect, and overly specific definitions were those that provided a correct definition of one kind of effect size, as if it were the only definition (for instance, writing that effect sizes are standardised measures, when in fact they do not have to be).

Once deductive coding was completed in full by EC (the thesis author), a random 20% subset of definitions was independently coded by RW (the primary supervisor). Cohen's kappa for inter-rater agreement was initially calculated to be 0.91 ('almost perfect agreement'; Landis & Koch, 1977), and after agreeing that three definitions had been mis-classified as 'acceptable' instead of 'shows understanding' by RW, kappa was revised upwards to 1.0 ('perfect agreement').

Once this deductive process had been carried out, a combined inductive-deductive approach was used to look for patterns within each category of *vague, overly specific,* and *incorrect* definitions in order to identify common misconceptions.

3.3.8 Participants

Participant demographics (*n = 247*) are shown in Table 3.2 and Table 3.3. The sample was quite evenly split between participants who identified as engaging with some aspect of open science (47%) versus those who did not (42.1%). PhD students made up nearly half of the sample (39.7%), with lecturers and postdoctoral researchers the second and third largest groups respectively (19.7% and 15.7%). Subsequent details about job position can be seen in Table 3.2. The largest country represented in the sample was the United Kingdom, which made up 66.5% of the sample with participants from all four respective nations (note that, as shown in Table 3.2, four participants reported their location as the United Kingdom instead of providing one of the four specific nations). The United States of America (USA) was the second most common location, making up 15.2% of the sample, with 14 other countries also represented, as shown in Table 3.2. A wide variety of sub-fields of psychology were also represented in the sample, the largest of which were health psychology (*n = 50*), cognitive psychology (*n = 23*), neuropsychology (*n = 22*) and social psychology (*n = 22*). The full spread of sub-fields is detailed in Table 3.3.

Table 3.2

Demographic Characteristics of the Sample (n =247)

Demographic Groups	Frequency
Job Role	
MSc Student	**16** (6.5%)
Research or Teaching Assistant (no PhD)	**10** (4.1%)
PhD Student or equivalent trainee	**98** (39.7%)
Postdoctoral Researcher	**36** (14.8%)
Lecturer or Senior Lecturer	**45** (18.2%)
Professor	**16** (6.5%)
Other[a]	**5** (2%)
Prefer not to say	**3** (1.2%)
Missing	***18** (7.3%)*
Location	
Australia	**5** (2%)
Belgium	**1** (0.4%)
Canada	**4** (1.6%)
Finland	**1** (0.4%)
France	**1** (0.4%)
Germany	**6** (2.4%)
Ireland	**6** (2.4%)
Italy	**3** (1.2%)
The Netherlands	**1** (0.4%)
New Zealand	**1** (0.4%)
Serbia	**2** (0.8%)
Singapore	**1** (0.4%)
Spain	**2** (0.8%)
Sweden	**3** (1.2%)
United Kingdom	**153** (61.9%)
England	***81** (32.8%)*
Northern Ireland	***2** (0.8%)*
Scotland	***63** (25.5%)*
Wales	***3** (1.2%)*
"UK"	***4** (1.6%)*
United States of America	**34** (13.8%)
Missing	***23** (9.3%)*
Open Science	
Yes	**116** (47%)
No	**104** (42.1%)
Prefer not to say	**9** (3.6%)
Missing	***18** (7.3%)*

[a] The category "other" includes one clinical psychologist, one occupational psychologist and part-time lecturer, one psychology practitioner, one research and training lead, and one participant who did not supply further details.

Table 3.3

Sub-Fields of Psychology in Sample

Field	Frequency
Affective	**3** (1.2%)
Applied	**3** (1.22)
Behavioural	**3** (1.2%)
Clinical	**19** (7.7%)
Cognition	**25** (10.1%)
Developmental	**10** (4.1%)
Educational	**2** (0.8%)
Environmental	**3** (1.2%)
Evolutionary	**4** (1.6%)
Experimental	**3** (1.2%)
Forensic	**16** (6.5%)
Health	**50** (20.2%)
Legal	**3** (1.2%)
Mental Health	**2** (0.8%)
Neuropsychology	**23** (9.3%)
Occupational	**3** (1.2%)
Personality	**3** (1.2%)
Psycholinguistics	**2** (0.8%)
Quantitative	**6** (2.4%)
Social	**22** (8.9%)
Sport	**4** (1.6%)
Other [a]	**8** (3.2%)
Missing	***30** (12.1%)*

[a] The 8 sub-fields classed as 'Other', each reported by one participant, were as follows: autism, comparative psychology, community psychology, counselling, cyberpsychology, decision science, methodology, and multidisciplinary psychology.

3.4 Results

In this sample, the majority of participants rated effect sizes as very important in psychological research; with no participants rating them as 'not important at all'. Further results are shown in Table 3.4. Two of the three participants who rated effect sizes as 'not very important' categorised themselves as engaging with open science in some way (1 x lecturer and 1 x professor), while the other participant did not (1 x postdoctoral researcher). Similarly, three of the six participants who opted for 'I don't know' also engaged with open science (1 x MSc student and 2 x PhD students), compared to three who did not (2 x PhD students and 1 x job role unknown).

Table 3.4

Perceptions of the Importance of Effect Sizes (n = 247).

Rating	Frequency
Very important	**168** (68%)
Somewhat important	**63** (25.5%)
Not very important	**3** (1.2%)
Not important at all	**0** -
I don't know	**6** (2.4%)
Missing	*7 (2.8%)*

3.4.1 Part 1: Using Effect Sizes

Effect size use was very high in this sample, with just 9.4% of participants indicating that they never use effect sizes in their work. It can be seen in Table 3.5 that within the small group of participants who do not use effect sizes, 70% do not identify as engaging with any aspects of the open science movement.

Table 3.5

Reported Effect Size Use, Split by Demographic Variables

Sample Group	Yes	Not Always	No
Sample	**175** (71.1%)	**48** (19.5%)	**23** (9.4%)
Open Science			
Yes	**95** (54.3%)	**17** (35.4%)	**4** (17.4%)
No	**65** (37.1%)	**23** (47.9%)	**16** (69.6%)
Prefer not to say	**5** (2.9%)	**3** (6.3%)	**1** (4.4%)
Missing	***10** (5.7%)*	***5** (10.4%)*	***2** (8.7%)*
Job			
MSc Student	**9** (5.1%)	**5** (10.4%)	**2** (8.7%)
Research or Teaching Assistant (no PhD)	**8** (4.6%)	0 -	**2** (8.7%)
PhD Student or equivalent trainee	**64** (36.6%)	**21** (43.8%)	**13** (56.5%)
Postdoctoral Researcher	**29** (16.6%)	**7** (14.6%)	**0** -
Lecturer or Senior Lecturer	**38** (21.7%)	**4** (8.3%)	**3** (13.0%)
Professor	**14** (8.0%)	**2** (4.2%)	**0** -
Other	**1** (0.6%)	**3** (6.3%)	**1** (4.4%)
Prefer not to say	**2** (1.1%)	**1** (2.1%)	**0** -
Missing	***10** (5.7%)*	***5** (10.4%)*	***2** (8.7%)*

A statistically significant difference was found between those who responded 'yes' and 'no' to the open science demographic question (χ^2 (2, N = 220) = 13.1, p = .001, V = .244). Looking at job roles, effect size use is noticeably high within the professor and postdoctoral

researcher groups, with no participants in these categories reporting zero use of effect sizes, although this did not correspond to a statistically significant difference (χ^2 (6, N = 221) = 11.4, p = .076, V = .16).

Not Using Effect Sizes

Content analysis identified four categories of self-reported explanations for not, or nor always, using effect sizes: lack of knowledge, lack of motivation, conditional use, and not doing quantitative research. All four categories were relevant to both *not using* and *not always using* effect sizes. These four categories and their associated sub-categories are listed in Table 3.6, with accompanying frequencies. Most commonly, participants justified not using effect sizes due to a lack of knowledge, or simply due to bad habits.

Table 3.6

Explanations for Not Using or Not Always Using Effect Sizes

Explanation	No	Not Always	Total
Lack of Knowledge			
Personal knowledge	3	8	**11**
Collective knowledge of statistics	2	5	**7**
Lack of Motivation			
Bad habit	4	7	**11**
Not required to use them	3	1	**4**
Only use if calculated by software automatically	-	1	**1**
Conditional Use			
Depends on audience	1	3	**4**
Depends on project	1	3	**4**
Educated preference	-	5	**5**
Other*	-	2	**2**
Not Doing Quantitative Research	4	1	**5**

As shown in Table 3.6, a lack of knowledge appears to exist on two levels in this sample: for some participants, it was clear that their own statistical knowledge was a barrier (e.g. *"I'm not totally sure of my statistical knowledge"*); while for others, a broader collective lack of knowledge related to particular statistics was an issue. For instance, one participant highlighted that there is *"debate over how to do this* [calculate effect sizes] *in linear regression (e.g. some people have told me beta is an effect size but others have said this isn't true)"*. Contrastingly, several participants indicated that a higher level of statistical knowledge influenced their behaviour, and that not using effect sizes was an educated

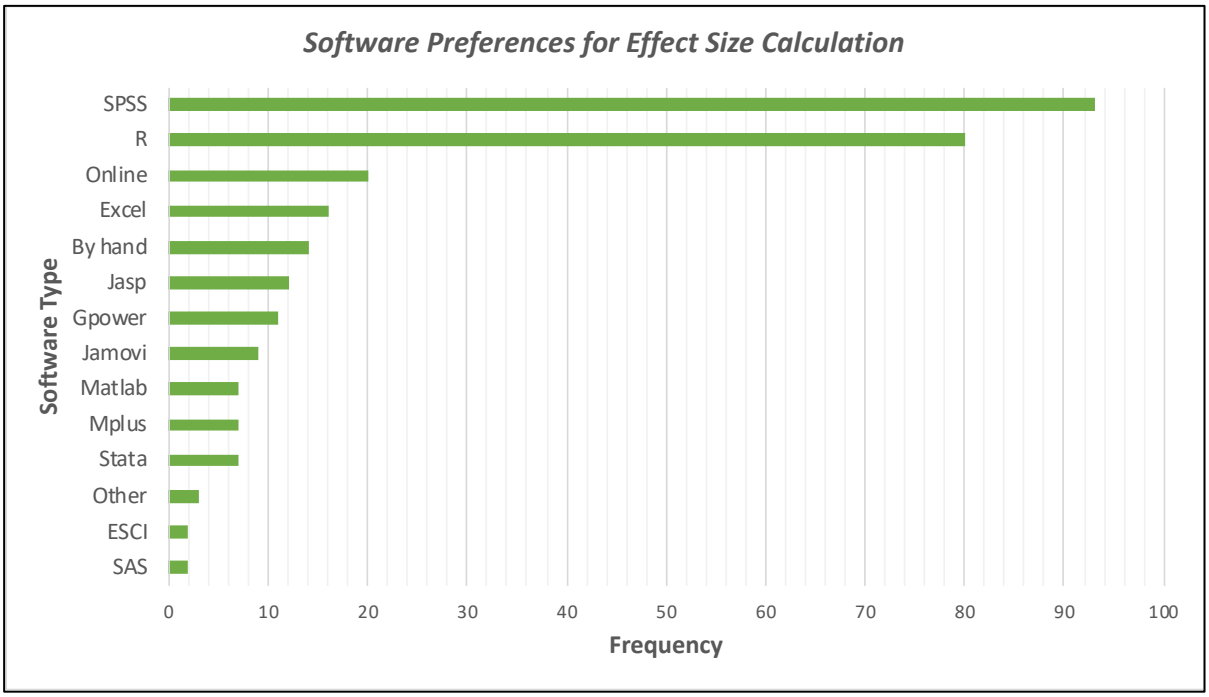
preference: for instance, *"sometimes they're not the most effective way to communicate the information and I instead use graphics and Bayes factors"*. The two explanations categorised as 'other' were only using effect sizes when directed by a supervisor, and only calculating effect sizes for significant *p*-values.

Software Preferences

When asked about their preferred software choices for effect size calculation, participants typically reported using multiple technologies. Figure 3.2 illustrates the variety of options listed across all responses, with the most frequent choices being SPSS© and R (mentioned 93 and 80 times respectively). The "other" category refers to three unique choices: BrawStats, Minitab and RevMan.

Figure 3.2

Software Choices Used by Participants for Effect Size Calculation



Publishing Effect Sizes

Experience including effect sizes in published papers was more varied than general reported effect size use. Just under one third of the sample (30.7%, or 74 participants) reported no experience of including effect sizes in published papers. Of the 69.3% (167 participants) who have included them in published work, 23 participants identified journal requirements as the motivation for incorporating effect sizes, 65 participants reported personal preference as their

motivation, and 77 participants said it was a combination of both factors. Two participants did not provide a response about preferences.

3.4.2 Part 2: True-False Testing

For each of the five statements, at least three quarters of participants chose the correct answer, with particularly high scores for Statement 4 and Statement 5 (although uncertainty was also high for Statement 5); the full spread of responses is shown in Table 3.7. Both of these statements (falsely) describe a specific relationship between effect sizes and NHST, but the majority of participants in this sample do not appear to hold any effect size fallacies. Twenty percent of participants incorrectly indicated that Statement 1 was false, while Statement 2 had the second-highest number of incorrect responses (16.2% mistakenly labelled S2 as true). Broadly speaking, knowledge related to effect sizes appears to be high when evaluated using this data.

Table 3.7

Response Frequencies For Each True-False Knowledge Statement (n = 234)

	Statement	True n (%)	False n (%)	I Don't Know n (%)
1	Effect sizes express the magnitude of the influence one variable has on another [a]	**176 (75.2%)**	45 (19.2%)	13 (5.6%)
2	A larger sample size will result in a larger (stronger) effect size	38 (16.2%)	**176 (75.2%)**	20 (8.5%)
3	If you are doing high quality research, your aim is to find the largest effect sizes possible	25 (10.7%)	**184 (78.6%)**	25 (10.7%)
4	A statistically significant *p*-value corresponds to a medium or large effect size	8 (3.4%)	**207 (88.5%)**	19 (8.1%)
5	A small effect size indicates that the null hypothesis should fail to be rejected	11 (4.7%)	**191 (81.6%)**	32 (13.7%)

[a] *Statement 1 was the only statement where "true" was the appropriate response.*

Note. Correct answers are highlighted using **bold underlined** text.

Only two participants scored zero out of five, one of whom was an MSc student and the other a PhD student. Both scores were a combination of incorrect and "I don't know" responses. Only five participants scored one out of five, including one lecturer. Twenty-five participants

scored two out of five, and 30 participants scored three out of five, all of whom ranged from students through to lecturers (a category which includes senior lecturers and equivalent staff members). Seventy-one participants scored four out of five, and 101 participants scored full marks.

A one-way ANOVA established that there was a statistically significant relationship between job role and true-false score, $F(7, 221) = 7.74$, $p < .001$, $n^2 = 0.197$. Mean and median scores for each job role are shown in Table 3.8, and it can be seen here that there was typically an increase in knowledge associated with more experienced job positions. In addition, a Welch's t-test identified a statistically significant difference between participants who responded yes or no to identifying with any part of the open science movement, $t(208) = 3.31$, $p = .001$, $d = 0.448$. Participants who responded *yes* had a mean true-false knowledge score of 4.27 (SD = 1.066), compared to participants who responded *no,* who had a mean score of 3.76 (SD = 1.195).

Table 3.8
Mean and Median True-False Knowledge Scores per Job Role

Job Role	Mean (SD) Score	Median Score
Professor	**4.69** (0.48)	5
Lecturer	**4.56** (0.66)	5
Postdoctoral Researcher	**4.17** (0.91)	4
PhD Student	**3.88** (1.20)	4
Research or Teaching Assistant	**3.8** (1.229)	4
MSc Student	**2.75** (1.29)	3
Other	**3.4** (1.52)	4
Prefer not to say	**2.33** (1.16)	3

3.4.3 Part 3: Defining 'Effect Size'

Participants were asked to define effect size in their own words. Just over one quarter of definitions were categorised as *acceptable*, more than half as *shows some understanding*, and the remaining 14.9% as *incorrect.* Further details are provided in Table 3.9.

Table 3.9
Categories and Frequencies for Definitions of Effect Size (n = 242)

Category	Frequency
Acceptable	**70** (28.9%)
Incorrect	**36** (14.9%)

Shows some understanding	**136** (56.2%)
Vague definitions	*56 (23.1%)*
Overly specific definitions	*80 (33.1%)*

***Acceptable* definitions**

Seventy definitions, or approximately 30% of the sample, were categorised as *acceptable.* Examples include "*the strength or degree to which one variable is related to another*" (Professor, psycholinguistics) and "*effect size measures the strength of the relationship between variables. different metrics exist, e.g. depending on the kind of variables involved*" (PhD student, personality psychology).

***Incorrect* definitions**

Thirty-six definitions were categorised as *incorrect*. Multiple participants confused effect sizes and other concepts, such as statistical power – e.g. "*[an effect size is] the power of your analysis*" – and degrees of freedom: "*an effect size is degrees of freedom*". Other participants incorrectly defined effect sizes as related to NHST, such as "*the value that adds meaningfulness to a p-value*", or the more jumbled "*how statistically significant something is in relation to participant numbers*". Some participants also demonstrated a misunderstanding of how effect size should be interpreted. For example, one participant added the following incorrect detail to their definition: "*if effect size is high, then you are more likely to be able to make reliable predictions than if it is low, in which case it is useless for making predictions and has little practical use as a result*". Participants who gave incorrect definitions also mentioned a variety of other ideas, such as effect size being a binary indicator of an intervention working or not, effect sizes being numbers which compare populations, and effect sizes being numbers which show how important an effect is. While the latter idea sounds plausible, it is important to remember that an effect size alone is simply a measure of magnitude: it may allow the importance of an effect to be evaluated, but does not do this on its own.

Definitions categorised as *shows some understanding*

Any definition that demonstrated an awareness of effect sizes but did not meet the criteria to be classed as *acceptable* was put into this category (*n = 136*). More than half of definitions were grouped into this category, which was then further divided into *vague* or *overly specific* definitions. Vague definitions were characterised by a lack of detail, typically offering short

descriptions such as "*effect size is the size of a statistical effect*", "*magnitude of a phenomenon*", or even more simply, "*measure of the effect*". The lack of detail in these responses meant that no further content analysis was plausible.

Overly specific **definitions**

Overly specific definitions indicated the broad misconception that effect sizes are limited to particular contexts, with several common more specific misconceptions occurring within this data. Note that many definitions included more than one misconception, and so frequencies within Table 3.10 add up to more than the total number of definitions classed as '*shows understanding – overly specific*' in Table 3.9. The specific common misconceptions found within this set of definitions were grouped into the three main categories: *effect sizes measure a difference, effect sizes measure experiments,* and *effect sizes are connected to statistical testing,* each with several subcategories. Two further common ideas were identified, along with four unique ones, all of which are shared in Table 3.10. These four unique ideas (categorised under 'Other' in Table 8) were: effect sizes are unstandardised values, effect sizes measure a real world effect, effect sizes provide meaning, and effect sizes measure clinical significance. The majority of these misconceptions indicate that knowledge of effect sizes may be constrained by personal experience or discipline.

Table 3.10

Overly Specific Definitions of Effect Sizes

Narrow Definitions	Frequency	Example Quote
Effect sizes measure a difference	**53**	
Difference (general)	8	*"the difference in parameters one is interested in"*
Group difference (general)	13	*"Magnitude of differences between groups"*
Two-group difference	20	*"Effect sizes reflect the magnitude of any given difference between two groups"*
Difference between conditions	8	*"Effect size means how much of a difference there is between conditions"*
Difference between means	4	*"a standardized estimate of the difference between your means"*
Effect sizes measure experiments	**11**	
Relates to experiments	4	*"A numerical measure of the scale/level of an effect observed in an experiment"*

Relates to interventions	2	*"The size of the change in the outcome following an intervention"*
Relates to manipulations	3	*"A measure of the strength of the effect of your manipulation on outcome variables"*
Relates to treatments	2	*"the proportion/percentage of variance due to the treatment"*
Effect sizes are connected to statistical testing	**10**	
Linked to inferential tests	8	*"The magnitude of the effect of a statistical test"*
Linked to significant *p*-values	2	*"When a significant association has been found between a dependant & independent variable, the effect size gives an indication of how much (how strongly) the independent variable changes the dependent variable"*
Further misconceptions	**21**	
Effect sizes are standardised metrics	15	*"tells us the magnitude of a reported effect using a standardized way"*
Effect sizes measure the strength of an interaction	2	*"Explains the strength of an interaction"*
Other	4	-

3.4.4 Part 4: Effect Size Training

More than half of the sample reported that they had not had sufficient training opportunities to learn about effect sizes, shown in Table 3.11. Despite 38.5% of the sample reporting that they *had* had sufficient training so far, only 7.9% of the sample responded 'no' when asked if they would be interested in further training, suggesting widespread interest in further education regarding statistics.

Table 3.11

Training Experience, and Interest in Future Training (n = 240)

Response	Frequency
Sufficient training so far	
Yes	**92** (38.3%)
No	**142** (59.2%)
Prefer not to say	**6** (2.5%)
Interest in further training	
Yes	**150** (62.5%)
Maybe	**71** (29.6%)
No	**19** (7.9%)

Eleven of the 19 participants not interested in any future training provided some explanation for their response, and all 11 attributed their lack of interest to having either sufficient knowledge or sufficient access to resources already. One of these participants indicated that instead of their own knowledge, they rely on the knowledge and resources of others, commenting "*I have enough statisticians in my teams that I don't need to worry about that [effect sizes] myself*".

Of the larger group of participants who were 'maybe' interested in training ($n = 71$), a further 34 also described themselves as having some level of sufficient knowledge already, but would be interested in context-specific or advanced training if it were available. Fourteen participants in the 'maybe' group highlighted that practical barriers were their concern with regards to training, both with regards to factors such as cost and location, but also in the context of their workload demand and a broader lack of time for self-development and learning.

3.5 Discussion

Effect sizes are a measure of magnitude which often provide researchers with useful evidence related to the research questions they are exploring, particularly in contrast to exclusively relying on NHST. While researchers such as Badenes-Ribera et al. (2016) have included questions about effect size in their wider statistics surveys, overall there is limited evidence related to effect size use and knowledge from the perspective of individual psychology researchers. The questionnaire study reported in this chapter makes a novel contribution to the literature by exploring effect size use and effect size knowledge in psychology researchers, using a combination of quantitative and qualitative data.

3.5.1 Using, or Not Using, Effect Sizes

Self-reported effect size use was high in this study, with fewer than 10% of participants indicating that they never use effect sizes, and a further 20% of participants 'not always' using them. These findings complement reviews of the literature, which have also found that effect size use is increasing over time (e.g. Peng et al., 2013). However, it should also be noted that the free-text definitions indicate that many researchers have a narrow perception of what an effect size actually is. In the context of effect size use, participants may be over-

estimating their own use based on a lack of knowledge about all of the appropriate circumstances in which various effect size indices can be used. For instance, if a researcher perceives an effect size to only measure differences (e.g. Cohen's *d*), they may identify as always using effect sizes in relevant circumstances, but in reality may not use them in other more diverse research contexts.

Self-reported explanations for not using effect sizes suggest that extrinsic motivation such as journal requirements are likely to positively influence behaviour, given that participants often attributed a lack of effect size use to 'not needing to use them' or more broadly explaining it away as a bad habit. These scenarios where researchers are choosing not to use effect sizes suggest that some researchers do not yet perceive the value of incorporating effect sizes into their work, signifying that intrinsic motivation is also a barrier to behaviour change. Given that the academic landscape has been dominated by a publishing culture which prizes $p < .05$ instead of carefully evaluated findings, this would not be surprising. This could also be connected to the narrow contexts in which many researchers in this study knew of effect sizes: if a researcher only knows that an effect size measures the size of an intervention, they may fail to see a need for effect sizes in their own work. Indeed, these explanations for not using effect size consistently reflect, both explicitly and indirectly, that a lack of knowledge is an important barrier to effect sizes.

3.5.2 Knowledge of Effect Sizes

A lack of knowledge was the most common self-reported explanation for not using effect sizes. This finding is reinforced in the training data, where fewer than 10% of participants reporting not being interested in effect size educational training. However, the true-false knowledge data *does* suggest there is at least a reasonable baseline of conceptual knowledge in this sample. For all five true-false statements, more than three quarters of participants correctly identified each as being true or false, which indicates that there is a good understanding of effect sizes within these topics, for most participants. Indeed, in this sample, very few participants appear to demonstrate the *magnitude fallacy* (a belief that there is a relationship between effect size magnitude and statistical significance), which has been identified in past studies (e.g. Kühberger et al., 2015).

However, a small but not negligible proportion (16%) of participants did indicate a belief that increasing sample sizes corresponds to increasing effect sizes, while a similar proportion of

researchers shared incorrect definitions of effect size. Several of these definitions demonstrate lingering confusion over effect sizes and NHST, such as those which described an effect size as a post hoc process to be calculated if a significant result has been found. This mirrors the uncertainty in responses to Statement 5 (‘*A small effect size indicates that the null hypothesis should fail to be rejected’*), where 14% of participants opted for ‘I don’t know’ as their response). Given that many other definitions of effect size indicated a partial, but not a comprehensive, understanding of effect size, coupled with a self-reported lack of knowledge and desire for better training, more progress needs to be made with regards to creating a solid foundation of new statistical knowledge in the psychology research population.

3.5.3 Study Limitations

It must be acknowledged that this sample does not generalise well to the wider psychology population. Firstly, the sample is skewed towards Western countries, with no African or South American representation, and only one participant in Asia (Singapore). Also, more than 65% of participants are early career researchers (in this context, meaning MSc research students through to postdoctoral researchers). In addition, a large proportion of the sample identified as being engaged with some aspect of open science or psychological reform, which is likely to mean more exposure to statistical discussions, and more interest in statistical reform. While it is impossible to establish the current proportion of psychological researchers worldwide who fall into this demographic group for the purpose of comparison, it is likely that the study sample over-represents this group, and that effect size use and knowledge could be lower in the wider population.

The measurements used in this online questionnaire should also be evaluated as a limitation of this study. The response options for use of effect sizes (*yes, not always, no*) fail to capture the finer details of effect size use, and would benefit from expansion to provide more options in future studies. Similarly, the statements used within the true-false measure are unvalidated and were devised as an exploratory tool for this study. They serve their purpose within this particular piece of research, but should not necessarily be used in future work without modification or expansion. In particular, the wording of the statements should be evaluated. The first statement in particular, which is used as a ‘correct’ definition of effect size, may mislead participants into implying causation (“*the magnitude of the influence one variable has on another*”), and so rating it as false may represent a critical perspective on causation, instead of a lack of knowledge. While the choice of statements here was grounded in the

literature, there are many other facets of effect size knowledge that could be explored. More broadly, the use of true-false statements also fails to examine knowledge in context, and so does not determine how well effect sizes are actually used and understood within research.

3.5.4 Future Directions

Future research which expands both the measures used here, and increases sample diversity and representativeness, is essential to build a more substantial picture of how effect sizes are used and understood in the psychology research population. As there is limited current research into effect size use and understanding, diverse study designs can be adopted to explore this area in a variety of ways. Data from these types of studies not only generally illustrates the knowledge levels of researchers, but also serves the purpose of providing clear directions for education and training that will improve statistical understanding and use. Chapter 4 of this thesis presents a further exploratory study related to effect sizes, using a more contextual research setting and new methods.

Outside of research, it must also be ensured that effect size reporting does not simply become a check-box exercise, with increased training and resources made available to researchers. While it is not the responsibility of journals to educate researchers, they do offer a particularly accessible way to reach a wide audience through their author guidelines, and could arguably easily provide helpful resources. This strategy has already been adopted by some journals, such as Psychological Science (as discussed in Chapter 2). Similarly, as the APA takes a clear stance on providing statistical instructions, they **c**ould expand their current selection of writing and formatting tutorials to also include statistical tutorials, given that they already have a big impact on how research is disseminated. It is clear based on the responses shared here that many psychology researchers are willing to take part in training and would engage with educational materials, if available. Unquestionably, individuals, universities and other institutions should also take responsibility for training, but a collective effort, utilising channels already widely accessed by researchers, is likely to have the most effective impact.

3.6 Conclusion: The Messy Reality of Effect Sizes

If effect size use is being strongly encouraged by organisations and editorial boards, then they must be used correctly, to avoid repeating past mistakes concerning statistics in psychology.

Therefore, the ideal psychology researcher must have a basic understanding of effect sizes, such as (but not limited to) the facts included in the true-false statements shared with participants in this chapter. In addition, the ideal researcher would have a deeper knowledge of effect sizes regarding their meaning and interpretation, and be able to use this deeper knowledge effectively and purposefully, to describe and compare the phenomenon they are studying. Without this, the *non-ideal* researcher is likely to encounter several issues, including (1) failing to report effect size indices in all relevant contexts, (2) making incorrect associations between effect sizes and NHST, and/or (3) not using effect sizes to evaluate data fully, i.e. thinking beyond 'there is an effect'.

The data presented in this chapter highlights the messy reality of effect size use and knowledge in a sample of psychology researchers. For instance, while scores were generally high on the true-false statements, up to a quarter of responses were incorrect per item, with uncertainty (measured using *I don't know* responses) as high as 13.7% for Statement 5 ('*A small effect size indicates that the null hypothesis should fail to be rejected')*. Perhaps the clearest example of this 'messy reality' is that effect sizes are only apparently understood in narrow contexts by many participants (e.g. *"a standardized estimate of the difference between your means"*); while many more are entirely unsure of what they are (e.g. "*how statistically significant something is"* or "*[an effect size is] the power of your analysis").* This lack of deeper knowledge will likely lead to effect sizes only being used in a small proportion of research settings, with researchers failing to make use of a wide range of indices to support their work. Additionally, researchers will struggle to accurately follow the advice of organisations such as the APA. It is clear that researchers need more information about the diversity of effect size indices, to expand their knowledge and to equip them with the correct statistics to support different styles of quantitative research.

Chapter 4: Exploring the Perception of Effect Sizes

Preface

This chapter presents a pilot study examining the perception of effect sizes on graphs. The original plan for this research was to collect data in a workshop focus group environment, to explore how researchers interpret statistics while also sharing their thoughts and questions. However, due to COVID-19, face to face research was not possible, and a pilot of an online workshop indicated that recreating this design online was not an effective way to collect data. Consequently, an online psychophysical approach was adopted instead, using a Method of Constant Stimuli-type design to present participants with a series of graphs. This experimental approach was inspired by 'night science', which encourages abstract approaches to explore "*the unstructured realm of possible hypotheses, of ideas not yet fully fleshed out*" (Yanai & Lercher, 2019). This exploratory approach presented the opportunity to study the uncertainty of effect sizes through a novel online experiment.

4.1 Abstract

Background: While conceptual knowledge is an important line of investigation, the understanding of effect sizes should also be examined in more contextual settings. At present, minimal research exists examining effect size understanding, and so this chapter uses an exploratory experiment to investigate the visual understanding of effect sizes.

Methods: An online experiment asked 56 UK-based psychology researchers to estimate the effect sizes shown on a series of graphs. Participants were randomised to view graphs displaying data from studies with $n = 50$ or $n = 100$ participants, and all participants saw a mix of scatter plots and two-group plots (a style of raw data presentation which would correspond to a t-test design). Effect size and statistical significance judgements were measured.

Results: Participants typically underestimated effect sizes throughout the experiment, with judgements of two-group plots slightly worse than scatter plots. However, participants who opted to use Pearson's r as their effect size index gave more accurate estimates than participants who used Cohen's d. Significance judgements were similarly inaccurate, and indicate that participants overestimate the critical effect size values which correspond to statistically significant results.

Conclusions: Overall, judgements of both effect size and statistical significance were poor. The slight advantage demonstrated by participants who used Pearson's r suggests that a stronger familiarity with correlations (and therefore a likely stronger familiarity with scatter plots) is an advantage for recognising effect sizes. Given that participants consistently overestimated the critical effect sizes for statistical significance, these findings suggest that a statistically significant result is imagined by researchers to correspond to a stronger effect than is actually the case.

4.2 Introduction

Thus far in this thesis, effect sizes have been examined through the lens of a straightforward questionnaire (presented in Chapter 3), which tested knowledge using a series of true-false statements and through qualitative definitions of effect size. However, this data does not provide a deeper insight into how effect sizes are perceived or understood. Given that an effect size is any value which quantifies the *size* of a phenomenon of interest, it would be useful to know how any particular effect size value is perceived by a researcher. This chapter presents the findings of a novel graph estimation study used to examine judgements of effect size.

4.2.1 Effect Size Judgement

Typically, effect sizes are written down, and the most popular guidance for interpretation is Cohen's benchmarks where $d = 0.2$ is described as a 'small' effect and so on (Cohen, 1992). However, both 'small effect' and '$d = 0.2$' are simply abstract descriptions, and very little is known about how these written descriptions correspond to an inner visualisation of a particular effect. As the primary value of an effect size is to provide deeper insight into the phenomenon being studied, and also to allow data to be compared, an ideal researcher should be able to form an accurate mental image of a given effect size in order to understand and make use of it. However, little is known about this thought process. In one of the few studies focusing on effect size perception, a sketch-the-effect-size method found that effect size judgements are highly inflated and widely varied (Kerns et al., 2020).

Even without explicit effect size indices, effects are often still of interest. For instance, even when relying exclusively on NHST and *p*-values, the majority of researchers are still studying an effect of some form, and are therefore likely to implicitly associate some degree of magnitude with it. However, little is known about the effect size judgements that researchers might subconsciously make when reading a result such as $p < .05$. In one of the only studies of statistical significance judgements using graphs, human judgement was found to be more conservative than actual statistical tests. Participants typically rated data as not being statistically significant, despite the data actually resulting in $p < .05$ when analysed with t-tests (the equivalent of participants making Type II errors in their judgements) (Sheth & Patel, 2015). This demonstrates that researchers expect to see a more obvious effect than is

required for data to be statistically significant. Conceptually, this aligns with the 'magnitude fallacy', which is where significant *p*-values are falsely associated with medium or large effects (Kline, 2004). Previous research has identified this in samples of psychology researchers (Oakes, 1986) and also psychology students (Kühberger et al., 2015); although just 3.4% of the participants sampled in Chapter 3 appeared to have this misconception.

4.2.2 Chapter 4 Overview

The study reported in this chapter was designed to explore judgements of effect size and statistical significance in an experimental setting, using a sample of psychology researchers. One option to measure judgement of both concepts would be to ask participants to produce their own graphs which correspond to particular effect sizes or *p*-values, to provide a clear insight into their individual perspectives – as used by Kerns et al. (2020). However, this is not a practical method, particularly when conducting research online. Instead, the study in this chapter adopted an approach similar to the traditional psychophysics "Method of Constant Stimuli", presenting participants with a range of stimuli to infer which ones represent particular effect sizes or *p*-values. Using an approach such as this makes it possible to measure both constant error (i.e. over- or under-estimation) and also variable error in judgements, which presents an insight into both perceptual bias and more general uncertainty.

Objective 1: To explore the accuracy of judgements of effect sizes that researchers make when inspecting graphical representations of data.

Objective 2: To explore researcher perceptions of statistical significance when inspecting graphical representations of data, to see if judgements of NHST (a more familiar and long-established concept) are better than judgements of effect size.

4.3 Methodology

This study was approved by the University of Stirling General University Ethics Panel (GUEP #1039 20-21), and adhered to the Code of Human Research Ethics guidelines of the British Psychological Society (BPS, 2014). Documentation can be found in Appendix C.

4.3.1 Sampling and Inclusion Criteria

As this study was designed as a small scale exploratory piece of work, no power analysis was used to determine a suggested sample size. An arbitrary target sample size of 50 responses was chosen, based on the availability of 5 x £20 Amazon vouchers as an incentive prize draw (to give participants an approximate 1 in 10 chance of winning a voucher). Similarly to the effect size survey study in Chapter 3, inclusion criteria for this study was any self-identifying psychology researcher, including PhD students. Location was restricted to participants within the United Kingdom (for the purposes of providing a prize in pounds sterling). Choosing to leave any questions blank did not impact entry into the prize draw. Participants were recruited via opportunity sampling across a two-week recruitment window in March 2021 through the internal Psychology department staff and PhD student mailing lists at the University of Stirling, and through Twitter.

4.3.2 Materials

An online questionnaire was created for this study, and delivered via Qualtrics (Qualtrics, Provo, UT). All materials are available within the OSF folder associated with this thesis (found here). The questionnaire was piloted with one lecturer and one PhD student, and based on their feedback the demographic questions were altered to exchange free-text boxes for a list of response items.

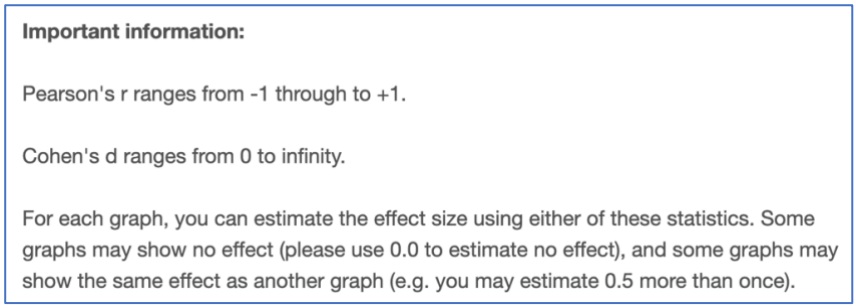
Participants were first presented with an information and consent sheet, and description of the task. This was followed with a brief series of demographic questions regarding job role and psychology sub-field. Participants were also asked about their engagement with any form of open science behaviours, SIPS or other elements of psychological reform using the same broad question reported in Chapter 3. Participants were then asked to rate their statistics knowledge using a sliding scale from 1 (very poor) to 10 (excellent).

The subsequent page asked participants if they calculate effect sizes in their own quantitative research, with a four-item response scale from *no – never* through to *yes – always*. This question was followed with a similar question about power analysis ("*Do you use a priori power analyses to determine sample sizes for all suitable quantitative research (i.e. not taking into account pilot/exploratory/qualitative work)?*"), with an added fifth response choice of *no – I have only conducted pilot, exploratory and/or qualitative work*. This information was collected to establish participant familiarity with effect sizes, both directly and also through power analysis calculations (which require effect size estimates). These

questions were followed with a reminder about effect sizes, as shown in Figure 4.1, to ensure that all participants were equipped with the same basic information about effect sizes before beginning the task.

Figure 4.1

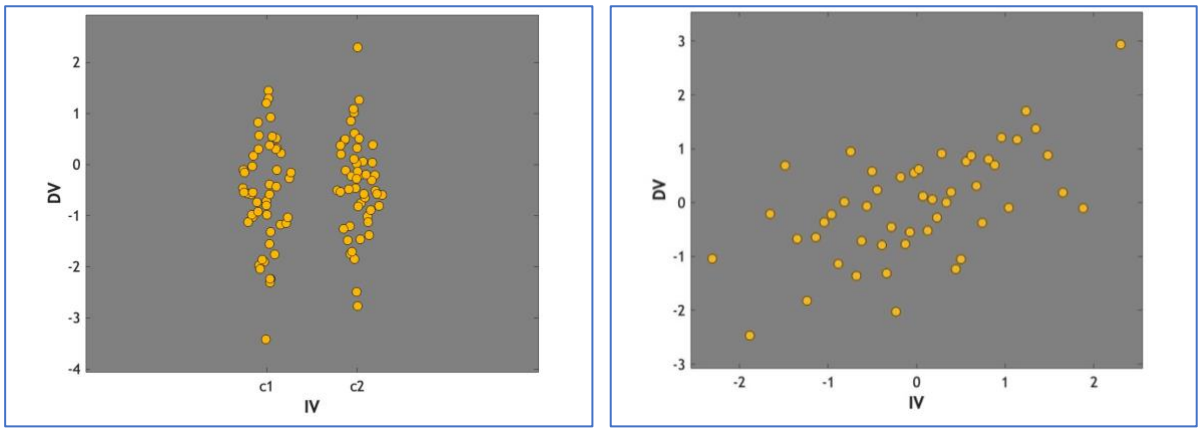
Information Shown to Participants Before Task



Eighty-four stimuli were created for this survey, split into two conditions. In both conditions, 21 scatter plots and 21 two-group plots were created, showing sample effect sizes ranging from $r = .0$ through to $r = .6$, in rising increments of 0.1. Each effect size was displayed on three different graphs, for each type of plot (e.g. each condition would include three different scatter plots showing $r = .4$, and three different two-group plots showing $r = .4$). All scatter plots were drawn from a bivariate normal distribution, and all two-group plots were drawn from a univariate normal distribution, using the same mean and standard deviation. Two example graphs are shown in Figure 4.2. The two conditions differed only by sample size: in Condition 1, graphs showed a sample size of $n = 50$, and in Condition 2, graphs showed a sample size of $n = 100$.

Figure 4.2

Sample Graphs Shown to Participants (Two-Group and Scatter Plots)



For each graph, participants were asked to make two judgements. They were first asked to estimate the effect size displayed (using a free-text response box), and to indicate whether they had used Cohen's *d* or Pearson's *r* for their estimate. This data is subsequently described as "Experiment 1". They were then also asked if they believed that the data would correspond to a statistically significant *p*-value ($p < .05$), with response options *yes, no* or *I don't know.* This data is subsequently described as "Experiment 2". Once all 42 stimuli had been presented, participants were asked to rate their confidence in doing the estimation task on a sliding scale from 1 (very poor) to 10 (excellent), and were asked whether they found one type of plot (two-group or scatter) easier to evaluate than the other.

4.3.3 Procedure

Participants gave informed consent via an electronic checkbox before beginning the experiment, and could leave by exiting the online webpage at any point with no penalty. Participants answered the demographic and background questions listed above, before being randomly assigned (by Qualtrics) to Condition 1 ($n = 100$) or Condition 2 ($n = 50$), making this a between-subjects design. Participants were then shown their 42 stimuli in a randomised order, and answered the questions listed above for each graph (covering Experiment 1 and Experiment 2 together), before answering the closing questions and entering the prize draw if they wished to do so.

4.3.4 Data Tidying

Sixty-seven participants began the survey. Eleven were removed from the final data set due to not continuing past the demographic questions, answering fewer than half of the estimated effect size questions, or using 0 or 1 for all of their estimated values. Once the final data set was assembled, all effect sizes reported using Cohen's *d* (as per participant preference) were converted to Pearson's *r*, using MATLAB, to create a cohesive data set. The conversion between Cohen's *d* and Pearson's *r* is as follows:

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

for example, as given in Ruscio (2008, p. 21). This equation is only exact when groups are of equal sizes, which is true of the data simulated for this experiment. Examples of Pearson's *r*

values when converted from Cohen's *d* are shown in Table 4.1, rounded to 4 decimal places (note this is not an exhaustive list of all estimates provided by participants).

Table 4.1

Conversion of Common Cohen's d Values to Pearson's r

Cohen's *d*	Pearson's *r*
0.2	0.0976
0.5	0.2357
0.8	0.3652
1.0	0.4472
1.5	0.6
2	0.7071

All effect sizes were then transformed using the Fisher z-transformation to regularise the distribution of *r* values, because the sampling distributions for z are normal and independent of effect size. In contrast, the sampling distribution for *r* varies in both standard deviation and skew (Fisher, 1925, as cited in Wicklin, 2017). Fisher's transform is given by:

$$z = tanh^{-1}(r)$$

Table 4.2.

Conversion of Pearson's r Using the Fisher Z-Transformation

Original	Transformed
0.2	0.2027
0.4	0.4236
0.6	0.6931

All effect size & significance analyses reported below were computed twice: once using data in the raw *r* format, and once using the Fisher z-transformation format. Results and patterns remained the same for both versions of the data. In order to provide graphs which best reflect the raw estimations provided by participants, the non-transformed values are used in the version of results presented here until the final comparison graph, which is clearly labelled in the results section of this chapter as using transformed values. This is further justified by the consideration of the set of effect sizes examined for this study: no graphs corresponded to an effect size greater than *r* = .6, and the transformed values remain very similar to the original values up to this point (see Table 4.2). Note that transformed graphs are available in Appendix C of this thesis for comparison.

4.3.5 Data Analysis

Descriptive summary statistics were computed for participant demographics, task confidence ratings, and graph feedback. In line with the pilot nature of this project, all subsequent data analyses were exploratory. Note that there is both a full series of results presented in Section 4.4, and a plain language summary (see 4.4.3).

The difference between each effect size estimation and actual effect size per trial was calculated. For each combination of sample effect size and plot type (e.g. scatter plot showing r = .1), participants gave three different estimates (as each combination was shown on 3 different graphs). These three responses were averaged to give a single estimate for each combination of effect size and plot type for each participant. Linear regression was then used to fit a straight line relating estimated effect size to actual effect size for each participant. The slope of this line indicates overestimation if > 1, underestimation if < 1, and perfect accuracy if equal to 1. The formula for this process is given by (where b is the slope):

$$r(estimated) = a + b * r(actual)$$

Data was then split by participant choice of effect size index (d or r) to look for potential differences in judgement between those who used Cohen's d and those who used Pearson's r.

Significance judgements were also analysed to identify how the proportion of *yes – this result would be statistically significant* ratings varied with the actual effect size for each graph type and condition. Exploratory probit analysis (a form of logistic regression; Finney, 1971) was used to fit a cumulative normal distribution to the significance data (shown in Figure 4.5 in Section 4.4.2). The effect-size at which the proportion of 'significant' responses crosses 50% was then estimated from the fitted curve for each participant. This is the value at which a participant was equally likely to rate a graph as significant or non-significant and is their estimate of the critical effect-size at which an effect becomes significant (hereafter referred to as their *perceived* critical effect size value). The *actual* critical effect size value (the effect size which would correspond to p = .05) for a sample size of 50 is r = .26, and for a sample size of 100 it is r = .19.

4.3.6 Participants

The final data set for this study consists of 56 participants, representing a wide variety of job positions and sub-fields of psychology. The demographic characteristics of this sample are reported in full in Table 4.3. Self-rated statistical knowledge was negatively skewed, with a mean and median of 6 (SD = 1.6) and mode of 7 out of 10, on a sliding scale.

Table 4.3

Demographic Characteristics of the Sample (n = 56)

Demographic Groups	Frequency
Job	
Research or Teaching Assistant (no PhD)	**1** (1.8%)
PhD Student	**21** (37.5%)
Demonstrator, Research or Teaching Assistant (with PhD)	**3** (5.4%)
Postdoctoral Researcher	**8** (14.3%)
Lecturer	**14** (25.0%)
Senior Lecturer	**3** (5.4%)
Professor or Tenured Staff	**3** (5.4%)
Reader	**1** (1.8%)
Other[a]	**2** (3.6%)
Field	
Clinical	**4** (7.1%)
Cognition	**18** (32.1%)
Cyberpsychology	**2** (3.6%)
Developmental	**2** (3.6%)
Educational	**1** (1.8%)
Evolutionary	**3** (5.4%)
Forensic	**2** (3.6%)
Health	**6** (10.7%)
Mathematical	**1** (1.8%)
Neuropsychology	**4** (7.1%)
Personality	**1** (1.8%)
Social	**5** (8.9%)
Other[b]	**5** (8.9%)
Missing	*2 (3.6%)*
Open Science	
Yes	**41** (73.2%)
No	**14** (25.0%)
Prefer not to say	**1** (1.8%)

[a] Other describes an MSc by research student and a public sector researcher with PhD

[b] Other sub fields: autism studies, comparative psychology, emotion studies, environmental psychology, and addiction

4.4 Results

Both effect size and power analysis use varied across this sample, with 10% of participants never using effect sizes, and 8.2% never using a priori power analyses, as illustrated in Table 4.4. Note that an a priori power analysis requires an estimated population effect size, and so is a reflection on effect size use in a wider context than just for accompanying results.

Table 4.4

Reported Effect Size and a Priori Power Analysis Use (n = 56)

Statistics Use	Frequency
Effect Size Use	
Never	**6** (10.7%)
Occasionally	**11** (19.6%)
Frequently	**19** (33.9%)
Always	**20** (35.7%)
Power Analysis Use	
Never	**5** 8.9%)
Occasionally	**16** (28.6%)
Frequently	**22** (39.3%)
Always	**9** (16.1%)
Not applicable to my work	**4** (7.1%)

The subsequent sections present participant judgements of effect size and statistical significance, illustrated with a series of graphs. These findings are then complemented with a plain language summary in Section 4.4.3.

4.4.1 Experiment 1 - Effect Size Judgements

Table 4.5 presents the spread of estimated effect sizes for each actual effect size, graph type, and sample size. The *n = 50* columns reflect data from participants in Condition 1 (small sample size), and *n = 100* reflect data from participants in Condition 2 (large sample size). On average, participants underestimated effect sizes regardless of graph type or sample size. However, estimates varied hugely, as demonstrated by both the standard deviations and the minimum and maximum estimated value: for instance, it can be seen that the maximum estimate for a graph showing $r = .0$ was $r = .8$. The most accurate estimates for each effect size are highlighted in **bold**, and show that typically (but not exclusively) estimates were more accurate for scatter plots and larger sample sizes.

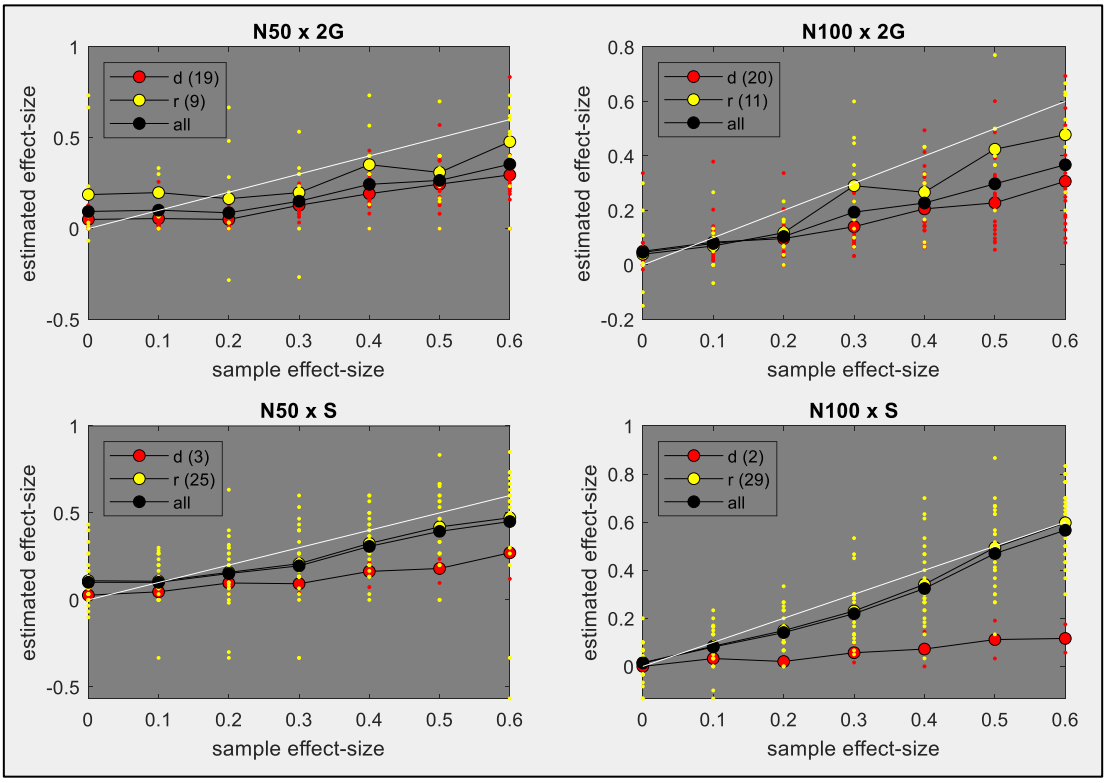
Table 4.5.
Collated Summary Statistics for Each Effect Size and Graph Type

Actual Effect Size	Estimated Effect Sizes			
	N = 50		***N = 100***	
	Scatter	**2-group**	**Scatter**	**2-group**
0.0				
Mean (SD)	0.084 (0.164)	0.079 (0.153)	**0.014 (0.094)**	0.046 (0.106)
Min	*-0.4*	*-0.1*	*-0.3*	*-0.25*
Max	*0.7*	*0.8*	*0.2*	*0.5*
0.1				
Mean (SD)	0.102 (0.186)	**0.099 (0.168)**	0.081 (0.127)	0.079 (0.111)
Min	*-1.0*	*0.0*	*-0.3*	*-0.2*
Max	*0.6*	*0.8*	*0.6*	*0.5*
0.2				
Mean (SD)	**0.166 (0.228)**	0.078 (0.192)	0.141 (0.128)	0.104 (0.106)
Min	*-1.0*	*-1.0*	*0.0*	*-0.2*
Max	*0.7*	*0.8*	*0.6*	*0.407*
0.3				
Mean (SD)	0.218 (0.242)	0.154 (0.198)	**0.219 (0.159)**	0.194 (0.151)
Min	*-1.0*	*-1.0*	*0.0*	*0.0*
Max	*0.7*	*0.8*	*0.7*	*0.7*
0.4				
Mean (SD)	0.308 (0.221)	0.260 (0.189)	**0.324 (0.202)**	0.227 (0.18)
Min	*0.0*	*0.0*	*0.0*	*-0.2*
Max	*0.8*	*0.816*	*0.9*	*0.816*
0.5				
Mean (SD)	0.393 (0.225)	0.291 (0.174)	**0.469 (0.217)**	0.282 (0.169)
Min	*0.0*	*0.0*	*0.0*	*0.01*
Max	*0.85*	*0.816*	*0.9*	*0.7*
0.6				
Mean (SD)	0.478 (0.326)	0.358 (0.184)	**0.566 (0.203)**	0.361 (0.192)
Min	*-1.0*	*0.098*	*0.05*	*0.049*
Max	*0.9*	*0.832*	*0.9*	*0.9*

Figure 4.3 provides more insight into the effect size estimates, divided by participant preferences for using Cohen's *d* or Pearson's *r*. Each graph plots the mean estimated effect size for each actual effect size, and the white slope indicates the expected estimate if participants were not making constant errors. In all four combinations of graph type and sample size, participants using Cohen's *d* (plotted in red) typically underestimated effect sizes to a greater degree than those participants opting for Pearson's *r* (plotted in yellow).

Figure 4.3

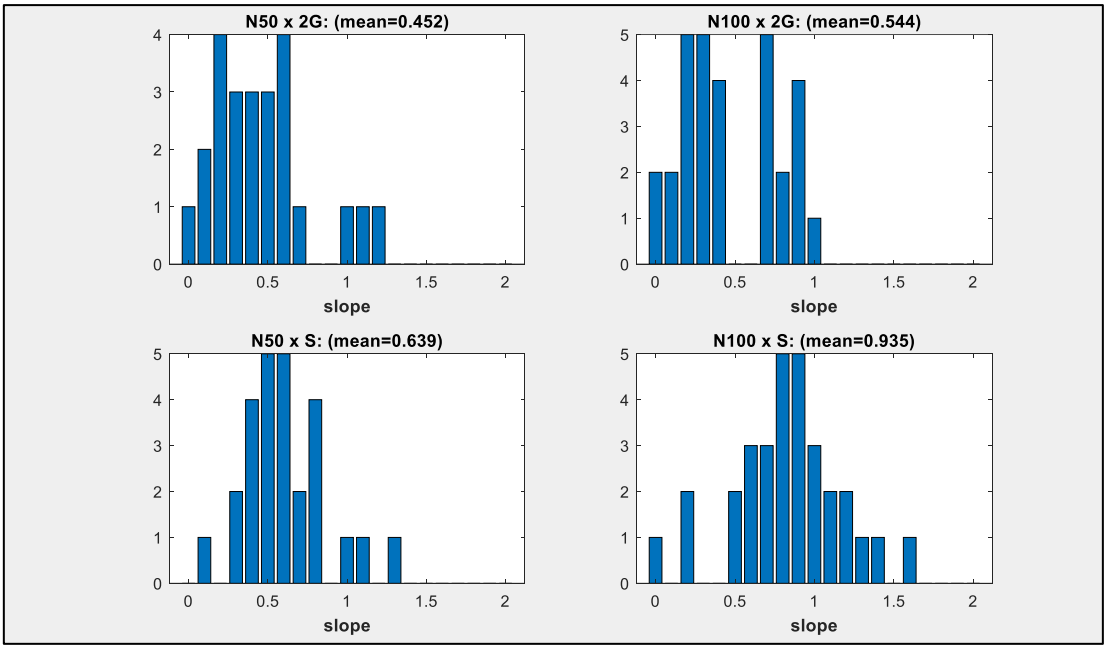
Effect Size Estimates for Each Graph Type and Sample Size



Note. Both axes are displayed as Pearson's *r*. 2G = two group plot and S = scatter plot.

Figure 4.4

Histograms Displaying Slopes for Individual Effect Size Estimates



Note. 2G = two group plot and S = scatter plot.

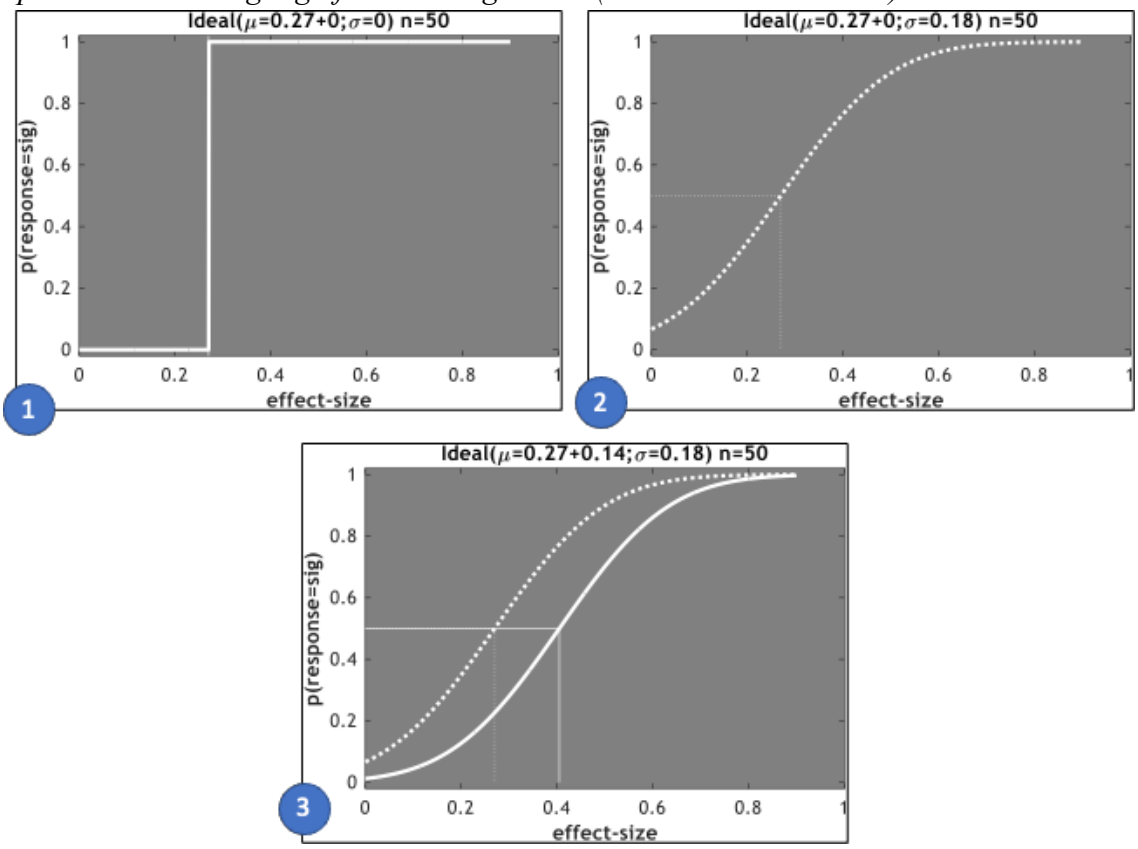
Figure 4.4 presents a different perspective on the effect size estimates, illustrating the varying participant accuracy slopes with a series of histograms. In the graphs shown in Figure 4.3, the white line corresponds to perfect accuracy, and has a slope value of 1. In Figure 4.4, it can be seen that participant slopes were only, on average, close to 1 in Condition 2 (n = 100), when assessing scatter plots (mean slope = 0.935). The highest degree of bias can be seen in both of the two-group conditions, even when sample size was doubled (n = 50 versus n = 100).

4.4.2 Experiment 2 – Significance Judgements

First, some context for this section is required. Judging the statistical significance of a data set on a graph requires two pieces of information: (1) the sample effect size and (2) the critical effect size value at which a data set within this design corresponds to a statistically significant p-value. An *ideal* participant would judge the effect size from the graph, and compare it with a known critical effect size. If the judged effect size is smaller than the known critical effect size, then the participant labels the graph as statistically non-significant. If it is larger, then they label the graph as statistically significant.

Figure 4.5

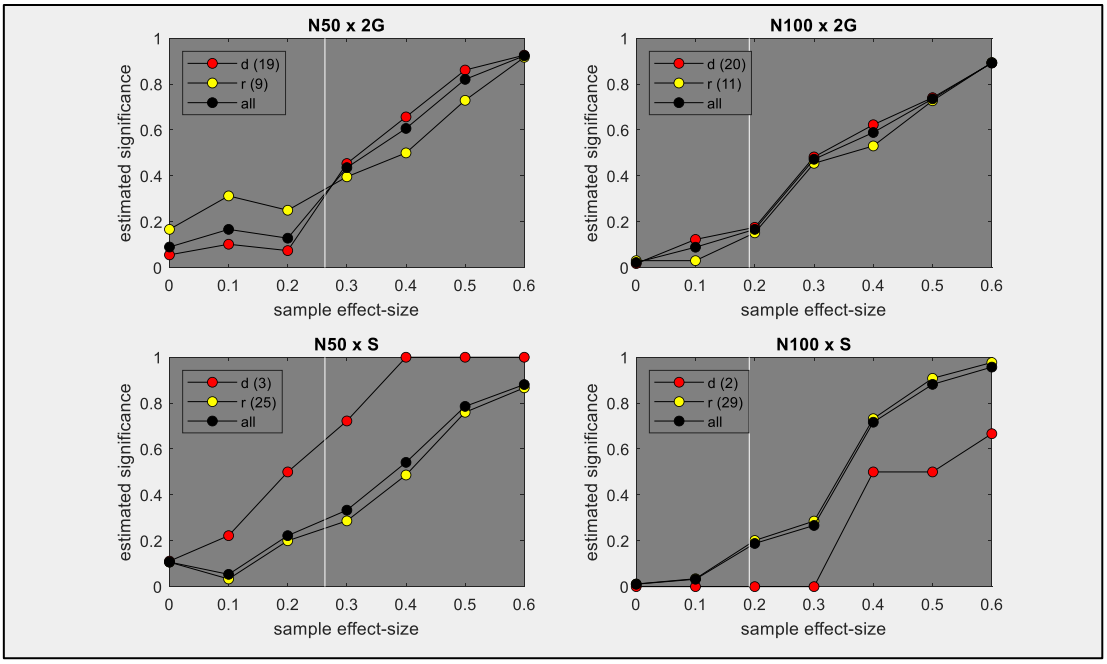
Graphs Demonstrating Significance Judgements (Ideal and With Errors)



Graph 1 in Figure 4.5 is a visual representation of an ideal set of results: there is a clear step from zero to one at the critical effect size, which corresponds to perfect judgements of statistical significance. There are three ways that the ideal participant's judgement then degrades. First, if the participant remains 'perfect' at knowing and using the critical effect size value, they could make random errors in judging the sample effect size. This creates a curve instead of a step from zero to one at the critical effect size line, as shown in Graph 2 within Figure 4.5 (note that the standard deviation of 0.18 is calculated from the effect size estimation data presented earlier in this chapter). The more variable the underlying effect size judgements, the flatter the curve. Second, with the same caveat of still perfectly using the critical effect size value, they could make a constant error in their judgement of sample effect sizes, which shifts the curve rightwards (for underestimation) or leftwards (for overestimation). Given that the data presented earlier in this chapter demonstrates constant underestimation, the curve has been shifted rightwards, shown in Graph 3 of Figure 4.5. Third, they may fail to use the critical effect size perfectly, and combine this mistake with errors of effect size judgement, which creates both joint random and joint constant errors. This has a further impact on flattening the curve, and shifting it rightwards or leftwards.

Figure 4.6

Significance Judgements for Each Graph Type and Sample Size

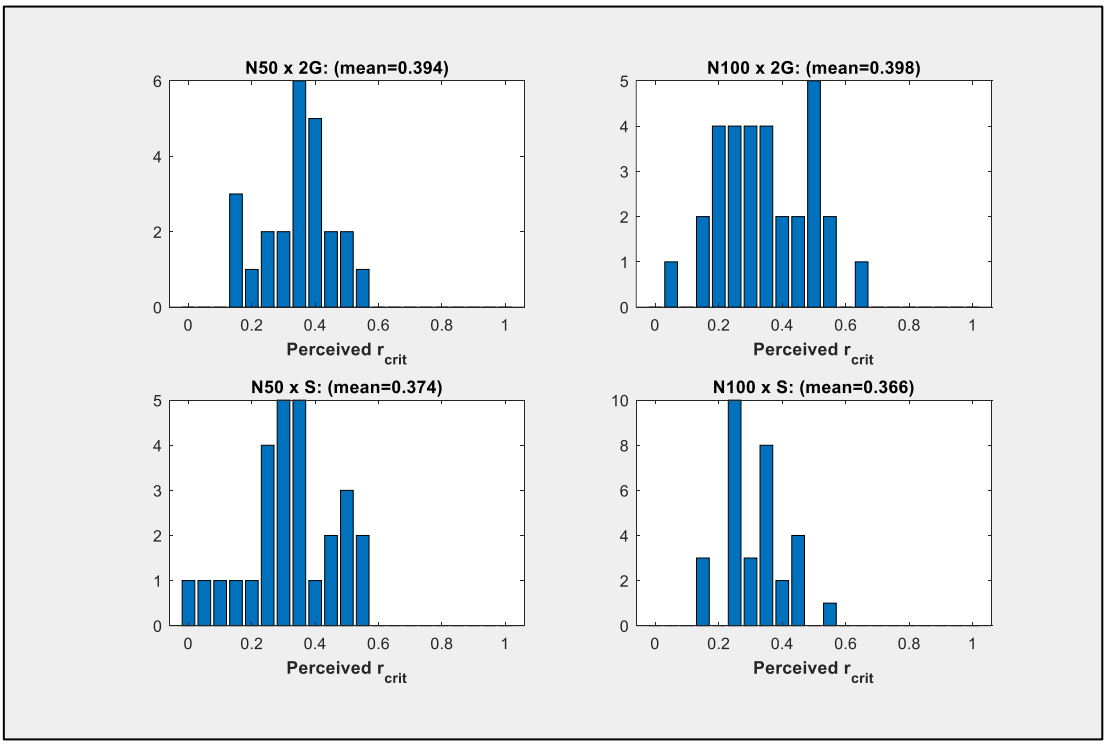


Note. The x-axis index is Pearson's *r* and y-axis measures proportion of significant judgements. 2G = two group plot and S = scatter plot.

Looking now to the actual results of this study: the proportion of 'significant' judgements for all participants combined, as a function of effect size, is shown in Figure 4.6 (once again divided into judgements made in Pearson's *r* and those made in Cohen's *d)*. The white lines on each graph illustrate the actual critical effect size value, i.e. the effect size where the graph type would correspond to $p = .05$ ($r = .26$ for $n = 50$ and $r = .19$ for $n = 100$). Recall that if all participants were rating the graphs correctly, the proportion on the left side of each white line should be zero, and proportions on the right side of the white line should be one. Across all four graphs, the curve is noticeably flat, and all four cross 0.5 on the y-axis to the right of the actual critical effect size, demonstrating that there is both high variable error and also high constant error in the judgements. While performance is broadly best in the N100 x S condition, there is wide variability across judgements overall. Despite a clear proportion of participants incorrectly rating very small effect sizes as significant, typically participants overestimated the effect size associated with statistical significance, which is explored further in Figure 4.7.

Figure 4.7

Perceived Critical Effect Size Values (r_{crit}) for Graph Type and Sample Size



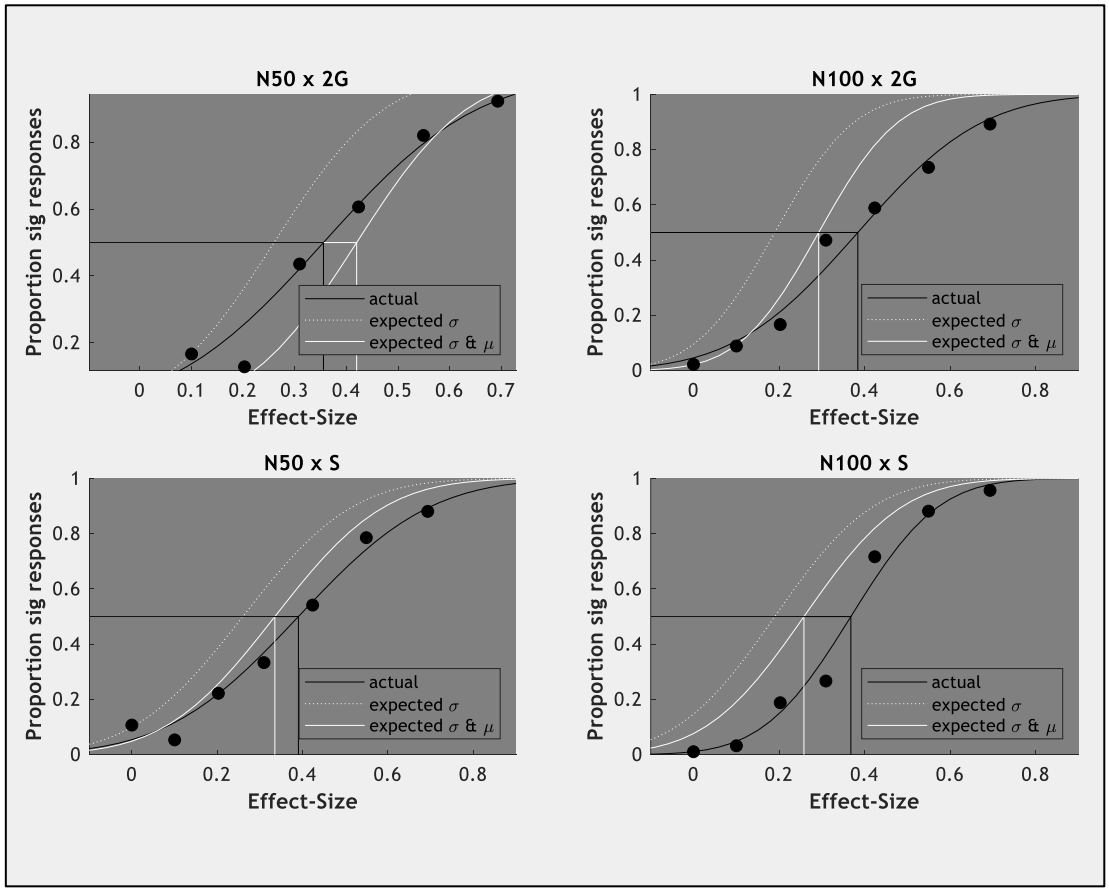
Note. 2G = two group plot and S = scatter plot

While each sample size has a fixed critical effect size, each participant also has a *perceived* critical effect size (the point at which a participant is equally likely to interpret a graph as

being significant or non-significant). Figure 4.7 presents the distribution of perceived critical effect sizes per condition. These graphs indicate that differences in these critical values vary only slightly between different graph types and conditions. The mean perceived critical values, ranging from 0.366 to 0.398, are all greater than the actual critical values for the graphs, (r = .26 for n = *50* and r = .19 for n = *100),* indicating that most participants overestimate the critical effect size associated with statistical significance. Noticeably, there are minimal differences between judgements made in the n = 50 condition versus the n = 100 condition.

Figure 4.8

Actual Versus Expected Significance Judgements



Note. Fisher's z-transformation has been applied to this data. 2G = two group plot and S = scatter plot.

The final set of graphs uses the responses from Experiment 1 (effect size estimation) to predict what responses to Experiment 2 (significance) should look like. The difference between what they are expected to be and what they actually are, as shown in Figure 4.8, is therefore the result of errors in perceived critical effect sizes. Recall from Figure 4.5 that ideal judgements would correspond to the perfect vertical 'step' from zero to one at the

critical effect size on the x-axis. Instead, the curves on the graphs in Figure 4.8 illustrate the actual judgements (in black), compared to the expected judgments which would be predicted using the random errors in sample effect size estimation (dotted white line) and random error + constant error from the overall underestimation of effect sizes (solid white line). The difference looks small visually, but there is more variability in the actual judgements versus the expected ones, suggesting that judgements of significance are made with a clear degree of randomness (arising from variability in the perceived critical effect size).

4.4.3 Simple Summary of Findings

Table 4.5 and Figures 4.3 – 4.8 present a variety of exploratory analyses of this data set. In Table 4.5, the mean estimates indicate consistent underestimation of effect sizes across all of the graphs, with large sample (Condition 2) scatter plots typically corresponding to the most accurate estimates. However, despite this, broadly there is no appreciable effect of sample size on accuracy (when considering all graphs as a whole). It is particularly important to acknowledge the huge range of effect sizes estimated by participants, such as the range of estimates from r = -.4 through to r = .7 for a graph showing an effect size of zero. Figure 4.3 offers a further insight into the difference between effect size index preferences, as Cohen's d is associated with bigger underestimations of effect size, while Figure 4.4 provides more evidence that the two-group plot style is associated with less accuracy in effect size estimation.

Figures 4.6, 4.7 and 4.8 demonstrate that significance judgements are also poor. In Figure 4.6, the flat curves on all four graphs illustrate high levels of variable error, and both this data and the data shown in Figure 4.7 suggest that participants typically have a larger critical effect size threshold for significance than exists in reality (which could indicate some belief that statistically significant results naturally correspond to more visible effect sizes). Finally, Figure 4.8 suggests that significance judgements are made almost at random, given that actual judgements are more variable than expected judgements.

4.4.4 Task Feedback

When asked whether they preferred a particular type of graph from the task, 65.6% of participants (n = 40) preferred the scatterplot, compared to just 13.1% (n = *8*) who preferred the two-group data graph. The remaining 21.3% of participants (*n = 13*) had no preference. With regards to task confidence, the mean self-rated score was 4.42 (SD = 1.63) out of 10,

where 10 is the highest possible rating. Responses to this question were slightly negatively skewed, with a median and mode of 5. There was no relationship between effect size estimation task accuracy and confidence, $r(54) = -.06$, $p = .65$.

4.5 Discussion

Chapter 3 established that researchers may have reasonable levels of basic knowledge about effect sizes, but typically only have a limited awareness of the variety of effect sizes that exist. The intention of the study presented in this chapter was to further explore effect size knowledge and understanding, to provide more insight into how effect sizes are understood by researchers. As there is broad scope for exploring effect size knowledge due to the current limited literature, and study designs were impacted by the COVID-19 pandemic, the study reported in this chapter makes a novel contribution to the literature by evaluating effect size perception with an experimental psychophysical design.

4.5.1 Findings and Implications

The findings of Experiment 1 highlight a clear disconnect between actual and perceived effect sizes, with actual effect sizes consistently underestimated by psychological researchers, and huge ranges of estimates made for a graph which corresponds to an effect size of zero. This suggests that when researchers read a result such as "$r = .3$", they mentally visualise a stronger effect than really exists within the data, reinforcing the findings of Kerns et al. (2020). Given that the most common range of effects within psychology is approximately $r = .1$ to $r = .3$ (Gignac & Szodorai, 2016), with most estimates placing the average effect size at $r = .21$ (e.g. Gignac & Szodorai, 2016; Fraley & Marks, 2007; Richard et al., 2003), the findings of this study suggest that researchers may view psychology as having much stronger effects than in reality. There appears to be a slightly advantage in using Pearson's *r*, given that the participants who preferred to use this index demonstrated slightly higher accuracy than those who chose to use Cohen's *d*. Participants opting to use Pearson's *r* may be more familiar with correlation analyses and therefore scatter plots, giving them an advantage when inspecting raw data as it may be more familiar to them.

Perceptions of statistical significance fared no better than effect sizes, despite the historic prevalence of NHST in psychological research. This long-running familiarity could be

expected to correspond to improved judgements, but this does not appear to be the case; although further research could compare early and late career researchers to test this specific hypothesis. Sheth and Patel (2015) found that researchers often make Type II errors when judging data, and similarly conservative judgements can be seen clearly here in Figure 4.6. In addition, the variety of errors and increased variability indicate that NHST appears to be used blindly, with no relationship between real data and *p*-values. Arguably, despite possibly being a more familiar statistic, these findings are not surprising given that NHST is often interpreted dichotomously with a lack of reflection on the meaning or uncertainty of a particular *p*-value. In comparison, the untidy and unclear nature of raw data does not allow for such straightforward decisions.

In both experiments, sample size also has no clear effect on judgement accuracy, despite participants in Condition 2 seeing twice as many data points as those in Condition 1. Given that there is currently a focus on increased statistical power (and therefore increased sample size), this offers the opportunity for future investigations related to whether increased sample size improves how data is perceived and understood.

4.5.2 Study Limitations and Future Directions

The study presented here is simply an exploratory pilot piece of work, providing a starting point for future investigations and hypothesis development. The sample size is small, heavily biased towards researchers who identify with some element of open science or psychological reform, and has few responses from participants at senior career stages (Senior Lecturer, Professor and similar). Hence, this data is not generalisable to the wider researcher population. If this experiment were to be repeated, it should be presented to much larger and more diverse samples, possibly with increased incentives to encourage wider participation.

With regards to the experiment itself, the number of trials must be acknowledged as a potential limitation. Exposure to a series of 42 similar stimuli may have led to decreased attention and therefore accuracy over time, which could have negatively influenced estimates. Despite this, each trial was very short and therefore not excessively cognitively demanding. In addition, self-rated task confidence had a median score of 5 out of 10, which suggests that the task sits in the middle ground between too easy and too difficult. In future research, attention checks at several intervals could be used to identify declining levels of attention or effort, perhaps offering participants drop-out points. In addition, despite finding that sample

size (n = 50 or n = 100) appears to have little effect on judgements, it is important to acknowledge that this variable can only be compared *between* participants. Future research should incorporate this into a within-subjects design, to more accurately compare judgement of smaller and larger data sets.

An improved version of this study should also utilise a platform which can provide feedback on estimation accuracy to participants at the end. Due to the two levels of randomisation (first to 1 of 2 conditions, and then to a randomly ordered series of graphs), this was not an available function in Qualtrics. Feedback would generate a useful educational opportunity for interested participants. Indeed, the findings here indicate that education in a broader sense would be valuable for researchers, whether that is focused specifically on effect sizes, or instead more generally supports researchers to visualise their data.

4.6 Conclusion: The Messy Reality of Effect Sizes (Part II)

As discussed in Chapter 3, the ideal psychology researcher would have both a basic knowledge of effect sizes, along with a deeper understanding which allows them to use effect sizes for interpretations and comparisons. The data in Chapter 3 provided a first insight into the messier reality of effect size knowledge, demonstrating that effect sizes may only be understood in very narrow contexts. The data in this chapter offers further tentative evidence for this messy reality, given that effect sizes were consistently underestimated by participants.

In addition, the data implies that significance judgements appear to be equally poor, despite researchers likely having much more exposure to NHST due to its prevalence in psychology. In comparison to these findings, the ideal researcher should be able to connect data and effect size values, and understand that p-values do not correspond to large or easily visible effects. This would enable them to accurately understand a set of findings. A more ideal reality may be one where researchers become accustomed to graphing data. Doing so is likely to help researchers understand that $p > .05$ is not necessarily a strong measure of evidence in favour of the effect they are studying. This makes the advantage of effect sizes more obvious, as the evidence can then be clearly quantified in a way that allows for it to be used and compared.

Chapter 5: The Use and Knowledge of Confidence Intervals in Psychology

5.1 Abstract

Background: Confidence intervals estimate a plausible range of values for a population parameter, based on a sample estimate (e.g. a sample mean), and provide an explicit measure of uncertainty. Reviews of the literature indicate that confidence intervals are used infrequently, and several studies suggest that confidence intervals are not well understood by researchers. This chapter offers new insights into confidence interval use and knowledge.

Methods: An online questionnaire was used to examine confidence interval use, self-reported barriers to use, and knowledge of confidence intervals in a sample of 206 psychology researchers. Knowledge was tested using a novel set of true-false statements, and by asking participants to define the term '95% confidence interval' in their own words.

Results: Just 10% of participants reported never calculating or reporting confidence intervals, while a further 41% only use them for some quantitative work. Notable self-reported barriers to use included a lack of knowledge, or being discouraged by supervisors or colleagues, while many participants also commented on only using confidence intervals when required to do so by journals. Overall, true-false knowledge scores were negatively skewed, with a mean of 2.43 out of 5 statements answered correctly. Based on the strict frequentist definition of a confidence interval, just 31 out of 206 participants correctly defined the term '*95% confidence interval'*.

Conclusions: In line with previous research, it appears that psychology researchers still hold a variety of misconceptions about confidence intervals. Participants also appeared to be divided over the correct interpretations of a confidence interval, with some participants indicating an awareness of the debates surrounding the strict definition of probability. Improved access to educational materials would support researchers in making well-informed decisions about their use and interpretations of confidence intervals.

5.2 Introduction

In Chapter 1, the estimation approach was introduced as one way to complement or replace NHST in psychology research. Confidence intervals are a key statistic within this approach, where they are used as an inferential statistic to question 'how certain is this finding?'. The review presented in Chapter 2 highlighted that many of the top 100 psychology journals now request that confidence intervals are included in submitted papers, and they are also strongly encouraged by organisations such as the APA. This chapter presents the findings of a second questionnaire study, which examines the use and knowledge of confidence intervals in a sample of psychology researchers.

5.2.1 What is a Confidence Interval?

As described in Chapter 1, a confidence interval is a range of plausible population values estimated from a measured sample parameter, typically used for the sample mean or sample effect size. Within the estimation approach, confidence intervals are intended to answer the question of 'how certain?', by inferring a range of possible population values from a single sample value. This is argued to be more valuable than exclusively relying on NHST for inferences, as *p*-values are typically only evaluated dichotomously: a single result simply is or is not statistically significant. Typically, this dichotomised thinking also neglects the possibility that any result may be an error (Type I or Type II), and is treated as certain evidence of an effect or no effect (e.g. Meehl, 1978; Hoekstra et al., 2006). In contrast, confidence intervals are designed to encourage critical evaluations of a finding.

A single interval lies inside the likelihood function of possible population values, avoiding the two tails of least likely values. Structurally, the interval presents upper and lower limits for likely population values, positioned around a measured sample value, such as 'M = 175 (SD = 20), 95% CI [168.8, 181.2]'. A confidence interval for a mean is constructed using the sample mean and sample size, and the z-distribution (for samples > 30), and uses the sample standard deviation as a best estimate of the true population standard deviation (when the true population SD is unknown). Note that any estimation of a population parameter from a sample either uses or implies some assumption about the distribution of population values (which applies to confidence intervals, as they estimate a range of population values).

Typically, in the absence of any further information, the expected distribution of all population means is assumed to be uniform; and so, the 95% confidence interval for an estimated population mean is the sample mean plus/minus 1.96 times the estimated standard error for the population mean. The formula for a 95% confidence interval is given by:

$$\left[norminv\left(\frac{(1-0.95)}{2}, mn_{\{samp\}}, \frac{sd_{\{samp\}}}{sqrt(n)}\right) \quad \ldots \quad norminv\left(1-\frac{(1-0.95)}{2}, mn_{\{samp\}}, \frac{sd_{\{samp\}}}{sqrt(n)}\right)\right]$$

For an effect-size, expressed as Pearson's *r* and assuming a uniform distribution of population *r* (the non-informative prior; a conventional assumption), then the 95% confidence interval is given by:

$$\left[tanh\left(norminv\left((1-0.95)/2, tanh^{-1}(r_{samp}), 1/\sqrt{n-3}\right)\right) \quad \ldots \quad tanh\left(norminv\left(1-(1-0.95)/2, tanh^{-1}(r_{samp}), 1/\sqrt{n-3}\right)\right)\right]$$

5.2.2 Use of Confidence Intervals in Psychology

Historically, confidence interval use has been low in psychological research. Several reviews of the 1990s literature found no evidence of confidence interval reporting across multiple journals (e.g. Keselman et al., 1998; Kieffer et al., 2001). Calls for their use have increased over time, particularly by the APA (Fidler, 2002; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Appelbaum et al., 2018), but this has not translated to consistently high use in psychology. For instance, despite being recommended in the 5th edition of the APA's Publication Manual in 2001 (Fidler, 2002), confidence intervals were identified in just 5% of articles published between 2002 and 2004 (Hoekstra et al., 2006). A subsequent meta-analysis of reviews estimated that, overall, confidence interval reporting only occurred in 10% of published articles in the period 1994-2006 (Fritz et al., 2013). However, more recent evidence suggests that the adoption of journal guidelines may positively influence reporting behaviour. For instance, an examination of the influence of journal guidelines at Psychological Science found that confidence intervals were reported in 70% of articles, rising from an earlier figure of 28%, after authors were 'strongly encouraged' to use the New Statistics (Giofrè et al., 2017). As shown in Chapter 2, 68 of the top 100 psychology journals now request or require confidence interval reporting for some or all quantitative work (as determined in October 2021), which is likely to correspond to increasing use of confidence intervals across more journals.

While several reviews have examined confidence interval use from a reporting perspective, very little research exists investigating their use from the perspective of individuals. In one of the only surveys which has examined self-reported confidence interval use, 6.4% of participants (in a study of $n = 472$ Spanish psychology researchers) never used confidence intervals, although just 26.1% of the sample reported using them 'quite often' (the top end of the scale given to participants) (Badenes-Ribera et al., 2016). In a subsequent replication (n = 159 Italian psychology researchers), despite nearly all participants reporting familiarity with confidence intervals, a slightly larger proportion (10.1%) reported never using confidence intervals, with a similar 27.7% using them 'quite often' (Badenes-Ribera et al., 2018). Neither survey asked for further data on confidence interval use or experiences.

5.2.3 Conflict About Confidence Intervals

Thus far, this chapter has described confidence intervals as if they are certainly a sensible addition or replacement for NHST, particularly given that they are recommended by major organisations such as the APA. However, while the premise of a confidence interval (a plausible range of population values) is uncomplicated, accurately interpreting the computed range of values is more complex.

Chapter 1 introduced the debate regarding the interpretation of confidence intervals, where researchers who align with a strict frequentist perspective disagree firmly with more flexible interpretations of a confidence interval. Strictly speaking, confidence intervals are defined and constructed in the context of long-run probability. This means that if a study was replicated 100 times, then 95% of the time, the study confidence interval would contain the true population value. This concept is not one that is debated: it is simply the mathematical property of a confidence interval. However, what this does mean is that one confidence interval alone either does, or does not, contain the true population value. Moreover, it can never truly be 'known' whether or not a single interval does or does not contain the true value (similar to the risk of a Type I error when using a p-value to reject a null hypothesis). Strictly speaking, interpreting a single confidence interval as '*having a 95% chance of containing the true population value*' is a *fundamental confidence fallacy* (Morey et al., 2016a), and is not mathematically correct.

However, other scholars argue that confidence intervals can be interpreted more leniently. While most researchers agree that the words 'chance' and 'probability' are too

mathematically-loaded to be used to interpret a single interval, researchers such as Cumming (2012, p. 79) argue that it is possible to be '95% *confident'* that a single interval contains the population value. As quoted in Chapter 1, Miller and Ulrich (2016) use a deck of cards analogy to make a similar argument for more lenient interpretations. It is broadly acceptable to discuss a deck of cards as if there is a 1 in 52 chance of a particular card being next in the deck, but under the strict definition of probability, the next card either is or is not a particular one (i.e. a 1 in 2 chance): perhaps the same degree of flexibility can be applied to confidence intervals. The work reported in this chapter integrates this conflict into the measures used and the interpretation of data, to examine which perspectives are currently held by researchers in psychology.

Key Definitions

Probability: The frequency at which a future confidence interval will contain the true population parameter. Once calculated, the single interval either has or has not captured it (although the outcome is unknown).

Confidence: A broad non-mathematical term used in place of 'probability' to indicate the strength of expectation that a single confidence interval, once calculated, is likely to contain a plausible range of potential population values, based on the available information.

5.2.4 Understanding Confidence Intervals

To ensure that reported confidence intervals are correctly interpreted and accurately calculated, researchers must be equipped with sufficient knowledge to understand what confidence intervals are, and how they should be understood. Past research has explored both basic and conceptual confidence interval knowledge across psychology. When examining very basic knowledge of confidence intervals, Fidler found that many students mistakenly perceived confidence intervals to be a range of plausible *sample* values, instead of a population statistic (Fidler, 2006), suggesting a lack of awareness of the inferential nature of confidence intervals. This could also be interpreted as a basic misunderstanding of the difference between samples and populations, which is more concerning as researchers with these misconceptions may not appreciate the sampling error which is present in nearly all research. Participants in Fidler's study also held misconceptions about confidence interval width, with almost three quarters of the sample incorrectly believing that a 90% confidence interval is wider than a 95% confidence interval. This lack of knowledge about confidence

interval widths has also been demonstrated in studies of PhD students, such as Kalinowski (2010).

The conceptual understanding of confidence intervals has been more widely studied, with evidence suggesting that researchers have a limited understanding of how they should be interpreted. For example, Cumming et al. (2004) demonstrated that many psychology researchers hold the *confidence-level misconception*, falsely believing that one 95% confidence interval captures 95% of means from future replication studies; when in fact, a single interval will (on average) contain 83.4% of future means (Cumming & Maillardet, 2006). In another early study, psychology researchers demonstrated similar difficulties with both basic and conceptual knowledge, such as confusing standard errors with confidence intervals, and also misunderstanding how confidence intervals can overlap when group differences are statistically significant (Belia et al., 2005).

Perhaps the most common approach to statistical knowledge testing has been the use of true-false statements, an approach initially used by Oakes (1986) and Haller and Krauss (2002) to test *p*-value knowledge, and also adopted in Chapter 3 of this thesis to study effect sizes. The process, which involves presenting participants with a series of all-false statements and asks them to label each as true or false, has been used to demonstrate that psychology researchers do not appear to understand confidence intervals. For instance, in a sample of 120 psychology researchers, only 3% of participants got a perfect score on a series of statements related to confidence intervals; with 11% of participants giving all-incorrect answers (Hoekstra et al., 2014). Within their study, nearly two thirds of the researchers held the belief that one interval has a 95% probability of containing the population mean (as critically discussed in Section 5.2.2), and similar numbers of participants believed that a confidence interval is a fixed interval which future means would fall into (the *confidence level misconception).*

This work has since been replicated, demonstrating similar levels of confidence interval misunderstanding in other samples such as Chinese psychology researchers (Lyu et al., 2018) and across other disciplines (Lyu et al., 2020). When evaluating this data, however, it should be noted that Hoekstra et al. (2014) argue fervently for the strict frequentist perspective of probability, and so consider statements such as "*we can be 95 % confident that the true mean lies between 0.1 and 0.4*" as false (one of the six statements used within these studies). They argue that this is evidence of "*a gross misunderstanding of CIs*" (p.1). However, from

Cumming's perspective, this is an acceptable way to interpret a confidence interval. As Hoekstra et al. (2014) and Lyu et al. (2018; 2020) did not ask their participants to provide more detail about their responses, it is impossible to know whether their findings represent true misunderstandings, or differing perspectives on probability.

A final recent study is that of Crooks et al. (2019), who developed a conceptual knowledge assessment for confidence intervals. Their study included examinations of misconceptions about confidence interval width and the sample misconception, similar to the work of Fidler (2006) and Kalinowski (2010), along with questions about sample variability, sample size, and interpreting confidence intervals. While their small sample mostly consisted of undergraduate students, the 19 graduate students in the study demonstrated high levels of misconceptions regarding replication and future means (the '*confidence-level misconception*'), and also typically held the belief that a confidence interval is an interval which '*you are 95% confident that the mean falls within'*.

5.2.5 Chapter 5 Overview

The intention of this study was to contribute new data on the use and knowledge of confidence intervals in the psychology researcher population, through a combination of quantitative and qualitative data. The specific objectives were as follows:

Objective 1: To examine confidence interval use including publishing habits and software preferences, and barriers to use, to capture individual perspectives and experiences.

Objective 2: To examine confidence interval knowledge using both a novel true-false statement test and through free-text definitions of the term '95% confidence interval', to identify knowledge levels and the prevalence and type of any current misconceptions.

Objective 3: To explore how researchers interpret confidence intervals, particularly comparing the strict mathematical interpretation related to long-run probability, versus the more flexible '95% confident' approach. Both quantitative and qualitative data contributes to this objective, with the intention of identifying how prevalent each interpretation currently is within this sample.

5.3 Methodology

5.3.1 Ethics

This study received ethical approval from the General University Ethics Panel (GUEP #857 19-20) at the University of Stirling, and adhered to the guidelines of the British Psychological Society (BPS, 2014). Documentation can be found in Appendix D.

5.3.2 Sampling and Inclusion Criteria

Due to the exploratory nature of this study, a power analysis was not deemed suitable for identifying a recommended sample size. Instead, opportunity sampling was once again used to recruit participants for this research, with the survey URL distributed on Twitter, and through academic mailing lists. Similarly to the questionnaire shared in Chapter 3, the study URL was distributed to the psychology staff and psychology PhD student mailing lists within the University of Stirling, and was also externally distributed to multiple JISCMail lists. Advertisements used for this followed the same structure as those used for the questionnaire study in Chapter 3 (see Appendix B for examples). The study reported here followed the same inclusion criteria as the effect size survey reported in Chapter 3: any psychology researchers actively involved in quantitative research in any location were eligible to take this study, including PhD and Masters-by-research students, but excluding undergraduate students.

5.3.3 Materials

Data was collected via an online questionnaire. All questions were developed by the researcher, incorporating statements and misconceptions shared in earlier research by others, which are detailed with references below. The questionnaire was piloted with one professor, one lecturer and one PhD student, with no changes recommended. The questionnaire can be found in Appendix D.

Define 95% Confidence Interval. Similarly to Crooks et al. (2019), participants were asked to provide a definition of the term *95% confidence interval* in their own words or write *I don't know* in the response box. This was asked as a free-text question to capture any misconceptions that researchers may have, and also to examine the prevalence of strict frequentist versus flexible definitions of a confidence interval.

Use of Confidence Intervals. Participants were asked if they currently calculate confidence intervals in their own quantitative research, with options *yes, no* or *not always*. If participants responded *no* or *not always*, they were asked to explain why not, if they were willing to do so. If participants responded *yes*, they were asked what software they used to compute their intervals. Participants were then asked if they had included any confidence intervals in any of their published papers or pre-prints, with options *yes* or *no*. Participants who responded *yes* were shown a follow up question, asking if they had done this due to *personal preferences, journal requirements, a combination of both*, or *prefer not to say*. This was asked to establish what may motivate psychologists to adopt confidence intervals in their research.

Perceived Importance of Confidence Intervals. Participant perception of confidence intervals was measured with the question "*how important do you feel confidence intervals are in psychological research?".* Participants answered using a Likert-style response item, with four options from *not important at all* (1) to *very important (*4*)*, with an alternative fifth option of *I don't know*.

Training in Confidence Interval Use. Participants were asked whether they felt that they had been provided with, or had access to, sufficient training on confidence intervals, with the response options *yes, no* and *prefer not to say*. They were then asked if they would make use of training, if it were made available, with options *yes, maybe, no* or *prefer not to say*. Participants who opted for *maybe* or *no* were shown a follow up open-ended question asking them to explain their response, if they wished to do so. Asking participants for their feedback was a way to identify potential barriers to engaging with training, which could then be tackled when offering future opportunities.

Scenarios. Participants were presented with two hypothetical scenarios sharing results in a confidence interval format, and were asked to give their interpretations. Further details of these scenarios, and the subsequent findings, are presented in Chapter 6.

True-False Knowledge Test. A six-item set of true-false statements was devised for this study, in a similar format to earlier investigations by Hoekstra et al. (2014), which are presented in Figure 5.1. However, instead of replicating Hoekstra et al.'s work, which focuses on interpretations of the word 'probability' and has been used by several researchers

(e.g. Lyu et al., 2018), this study compiled an assortment of statements based on the misconceptions of confidence intervals from previous research.

Figure 5.1

Hoekstra et al. (2014)'s Six True-False Statements

Statement
1. The probability that the true mean is greater than 0 is at least 95%
2. The probability that the true mean equals 0 is smaller than 5%
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect
4. There is a 95% probability that the true mean lies between 0.1 and 0.4
5. We can be 95% confident that the true mean lies between 0.1 and 0.4
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4

The statements chosen for this study represent a broader range of facts about confidence intervals including width, samples versus populations, replication, and interpretation. Details of the statements and their sources are in Table 5.1.

Table 5.1

Six True-False Statements Presented to Participants

	Statement		**Source:**
1	A 95% confidence interval is the range of values for which you are 95% confident that the population mean falls within.	**T/F**	Cumming, 2012; Miller & Ulrich, 2016
2	If all other factors are held constant, an 80% confidence interval will be wider than a 95% confidence interval.	**F**	Crooks et al., 2019
3	If all other factors are held constant, a confidence interval from a sample of n=25 will be wider than a confidence interval from a sample of n=100.	**T**	Crooks et al., 2019
4	A confidence interval gives you the range of plausible values for the true sample mean.	**F**	*based on* Fidler, 2006
5	If an experiment is replicated with new samples from the same population, 95% of future means will fall within the original 95% confidence interval.	**F**	Cumming et al., 2004
6	If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean	**T**	Morey et al., 2016

Note: F = False And T = True. "I don't know" was also an option for each statement.

It is important to note that these are novel statements, which have been combined into an unvalidated scale. As this is exploratory research, validation was not deemed necessary in the context of this single study (although a reliability analysis and correlation matrix can be found in Appendix D). Scale development, statement choices and potential future use will be evaluated later in this chapter.

In Hoekstra's study, participants were presented with five all-false statements, which has been criticised for potentially misleading participants (Miller & Ulrich, 2016). In the same manner as the true-false statements used in Chapter 3, the statements in this study were presented alongside a third "I don't know" option to minimise guesswork, and both true and false statements were included so as to not deceive participants into second-guessing their responses. In this study, statement 1 (Table 5.1) would be considered false under the strict frequentist perspective argued by Hoekstra et al. (2014), but would be considered true in accordance with the flexible interpretation favoured by Cumming (2012) or Miller & Ulrich (2016).

Demographics. Limited demographics were collected for this survey, in the same style as the effect sizes survey reported in Chapter 3. Once again, participants were asked for their location, field of psychology, and job role, and engagement in any kind of psychological reform. This question was phrased as "*are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar?",* with response options yes, *no* or *prefer not to say*. Throughout this chapter, this variable is abbreviated to "Open Science" for brevity.

5.3.4 Procedure

Qualtrics (Qualtrics, Provo, UT) was used to deliver the questionnaire online. The questionnaire began with an information sheet and digital consent form, with the eligibility criteria explained within the sheet and then listed as a checklist within the consent section to confirm participants were suitable to take part. The questions were shown in the order listed in Materials, and it was signposted that all questions were optional and could be left blank if

preferred. Participants could exit the questionnaire at any point with no penalty, and had seven days to return to the URL and finish their entry, if they wished to do so.

5.3.5 Data Handling

Overall, 235 participants began the survey, but 29 participants were removed from the data set for one of the following reasons: not progressing beyond the consent page, not being eligible due to being an undergraduate student, or writing "I don't know" or other nonsensical responses throughout the survey. The final data set analysed in this chapter consists of data from 206 participants, details of which are provided in section 5.3.8.

5.3.6 Quantitative Analysis

Quantitative data was analysed using Jamovi (The Jamovi Project, 2021), with descriptive statistics computed for all quantitative questions. Demographic differences were also explored using chi square tests, t-tests, or one-way ANOVAs as appropriate. Note that the variable Job Role was collapsed into four categories (pre doctoral level participants, doctoral students, post-doctoral academics and tenure-level staff) for the purpose of chi-square analyses, due to small group sizes. Exploratory 'knowledge' scores were computed per participant based on the number of correct answers they gave to the true-false scale test. These scores were calculated using responses to statements 2-6, as statement 1 could be marked as true or false depending on perspective (as discussed in Section 5.2.2).

5.3.7 Qualitative Analysis

Qualitative free-text responses explaining reasons for not using confidence intervals and for not being interested in training were analysed using a basic content analysis, taking an inductive (bottom-up) approach to code and categorise responses (Drisko & Maschi, 2015). Frequencies were then counted to identify the most common explanations shared by participants. Note that the process of basic content analysis is described in more detail in Chapter 3, Section 3.3.6.

Content Analysis: Definitions of '95% Confidence Interval'

Similarly to the approach used in Chapter 3 to analyse definitions of the term 'effect size', content analysis was used both deductively and inductively to examine definitions of the term '95% confidence interval'. First, definitions were coded deductively (a top-down approach) as either *incorrect* or *correct*. This deductive process was carried out twice: first, in

accordance with the strict frequentist approach advocated by researchers such as Morey et al. (2016a), and then once more using the more flexible approach adopted by researchers such as Cumming (2012) and Miller & Ulrich (2016). An example of this is shown in Figure 5.2. Recall that having 95% confidence that one interval contains the population mean is a misconception called the *fundamental confidence fallacy* by Morey et al. (2016a) but is considered plausible by Cumming (2012).

Figure 5.2

Example of Deductive Coding Using Strict and Flexible Approaches

Definition	Strict Approach	Flexible Approach
You are 95% confident that the true (population) parameter lies within that interval.	Incorrect	Correct

The full data set of definitions was deductively analysed by Rater 1 (EC), with a random 20% subset deductively analysed by Rater 2 (RW) for the purposes of inter-rater reliability. When rating definitions using the strict approach, agreement as measured by Cohen's kappa was 1.0 (indicating 'perfect agreement'; Landis & Koch, 1977). When rating definitions using the flexible approach, agreement was initially measured as 0.92 ('almost perfect agreement'), and was revised upwards to 1.0 after agreeing that a small number of unclearly worded definitions should be classed as 'incorrect'. Once categorisation was complete, definitions rated as incorrect under both the strict and flexible approaches were coded inductively by EC to identify any common misconceptions that were shared by participants.

5.3.8 Participants

The final data set is made up of 206 participants from an assortment of sub-fields of psychology, countries, and job roles; shown in Table 5.2 and Table 5.3. Just over half of the participants were from the United Kingdom (note that five participants did not provide details of their particular nation within this data), with the United States of America being the second-most common reported location. Half of the participants self-identified as engaging with some aspect of psychological reform or open science, as indicated by the Open Science variable.

Table 5.2

Demographic Characteristics of the Sample (n = 206)

Demographic Groups	Frequency
Job Role	
MSc Student	**10** (4.9%)
Research or Teaching Assistant (no PhD)	**2** (1%)
PhD Student or equivalent trainee	**64** (31.1%)
Postdoctoral Researcher	**30** (14.6%)
Lecturer or Senior Lecturer	**47** (22.8%)
Professor	**13** (6.3%)
Other[a]	**3** (1.5%)
Missing	***37** (18%)*
Location	
Australia	**1** (0.5%)
Austria	**1** (0.5%)
Canada	**4** (1.9%)
Cyprus	**1** (0.5%)
Finland	**1** (0.5%)
Germany	**7** (3.4%)
Republic of Ireland	**4** (1.9%)
Israel	**1** (0.5%)
Italy	**1** (0.5%)
Poland	**1** (0.5%)
Portugal	**1** (0.5%)
South Africa	**1** (0.5%)
Sweden	**2** (1%)
The Netherlands	**2** (1%)
United Kingdom	**113** (54.9%)
England	***60** (29.1%)*
NI	***2** (1%)*
Scotland	***43** (20.9%)*
Wales	***3** (1.5%)*
"UK"	***5** (2.4%)*
United States of America	**21** (10.2%)
Prefer not to say	**1** (0.5%)
Missing	***43** (20.9%)*
Open Science	
Yes	**103** (50%)
No	**66** (32%)
Missing	***37** (18%)*

[a] Three 'other' jobs in this sample were: industry psychology researcher, clinical psychologist, and occupation not shared.

Many sub-fields of psychology were represented in this sample. The most common were cognition, health psychology and social psychology, making up 15%, 12.6% and a further 12.6% of the sample respectively. The 'Other' category in Table 5.3 spans a wide range of other sub-fields, including cyberpsychology, metascience, and music psychology.

Table 5.3
Sub-Fields of Psychology in Sample (n=206).

Field	Frequency
Clinical	**9** (4.4%)
Cognition	**31** (15%)
Counselling	**2** (1%)
Cross-Cultural	**2** (1%)
Developmental	**13** (6.3%)
Educational	**3** (1.5%)
Evolutionary	**2** (1%)
Experimental	**2** (1%)
Health	**26** (12.6%)
Mathematical	**3** (1.5%)
Neuropsychology	**19** (9.2%)
Organisational	**2** (1%)
Personality	**3** (1.5%)
Psycholinguistics	**2** (1%)
Research Methods	**3** (1.5%)
Social	**26** (12.6%)
Sports	**2** (1%)
Other [a]	**10** (4.9%)
Missing	***46** (22.3%)*

[a] The 10 participants classed as 'Other' identified as: biopsychology (1), comparative (1), cyber (1), environmental (1), forensic (1), interdisciplinary (1), metascience (1), music (1), and unknown 'research psychology' (2).

5.4 Results

When asked about the importance of confidence intervals in psychology, 83% of participants rated them as either *very important* or *somewhat important*, as shown in Table 5.4. Perceptions of the importance of confidence intervals did not differ by job role (χ^2 (166, 9) = 3.38, p = 0.948, V = 0.08) or by open science category (χ^2 (169, 3) = 4.30, p = .231, V = 0.16).

Table 5.4

The Importance of Confidence Intervals in Psychological Research

Importance	Frequency
Very important	**94** (45.6%)
Somewhat important	**77** (37.4%)
Not very important	**12** (5.8%)
Not important at all	0 -
I don't know	**8** (3.9%)
Missing	**15** (7.3%)

5.4.1 Part 1: Using Confidence Intervals

Most participants in this sample reported using confidence intervals for all or some of their quantitative research, with just 10% never using confidence intervals at all. Use of confidence intervals did not differ statistically significantly by job role (χ^2 (165, 6) = 2.87, p = 0.825, V = 0.09, or between participants who do and do not engage with open science in some way (χ^2 (168, 2) = 2.07, p = .355, V = 0.111); a full demographic breakdown is shown in Table 5.5.

Table 5.5

Use of Confidence Intervals, With Demographic Differences (n = 203)

	Use of Confidence Intervals		
	Yes	*Not Always*	*No*
Full Sample	**100** (49.3%)	**83** (40.9%)	**20** (9.9%)
Job Role			
MSc Student	**6** (6%)	**3** (3.6%)	**1** (5%)
Research or Teaching Assistant	**1** (1%)	**1** (1.2%)	**0** -
PhD Student	**31** (31%)	**25** (30.1%)	**7** (35%)
Postdoctoral Researcher	**17** (17%)	**11** (13.3%)	**2** (10%)
Lecturer	**27** (27%)	**17** (20.5%)	**3** (15%)
Professor	**5** (5%)	**7** (8.4%)	**1** (5%)
Other	**2** (2%)	**1** (1.2%)	**0** -
Prefer not to say/missing	***11** (11%)*	***18** (21.7%)*	***6** (30%)*
Open Science			
Yes	**56** (56%)	**40** (48.2%)	**6** (30%)
No	**33** (33%)	**25** (30.1%)	**8** (40%)
Prefer not to say/missing	***11** (11%)*	***18** (21.7%)*	***6** (30%)*

Just 20 participants reported never using confidence intervals in their work, with a further 83 participants indicating that they did not always use them. Table 5.6 presents all of the explanations given for not using, or not always using confidence intervals. Reasons given included lacking knowledge about confidence intervals and being influenced by supervisors

and the behaviour of others more broadly, such as "*It's not requested by my boss - I'm a post doc"* and "*[I don't do it because] often the literature I am using doesn't include them".* Many participants also acknowledged only using confidence intervals when required to do so by a journal. Other participants offered more informed perspectives such as preferring to use Bayesian statistics, or being generally critical of confidence intervals. Critical participants typically invoked the strict frequentist approach to confidence intervals to support their choice, for example: "*because it provides only a range of possible values not the true value or how good your estimate of the true value is (precision fallacy) because you do not know if your 95% CI is one of the 95% that contains the true value or one of the 5% that does not"*.

Table 5.6

Reasons for not Calculating Confidence Intervals

Reason	**Use of Confidence Intervals**	
	No	*Not Always*
Depends on journal requirements	**2**	**11**
Lack of manuscript space or word count	**-**	**5**
Prefer other statistics such as Bayesian	**4**	**7**
Depends on research or analysis	**2**	**17**
Influenced by others	**2**	**5**
Lack of knowledge or experience	**3**	**14**
Bad habit	**2**	**5**
Use when calculated by software automatically	**-**	**3**
Critical of CIs	**3**	**4**
Other *(only use when included on a graph)*	**-**	**1**

Of the 183 participants who do calculate confidence intervals (those who responded *yes* or *not always* in Table 5.5), 123 confirmed that they have included them in a published paper or pre-print. When asked to think back to why they included confidence intervals for the *first* time in a paper, only 35 (28.5%) of these participants reported including them due to journal requirements, compared to 88 (71.5%) saying that they were initially motivated by personal preferences.

With regards to software use, SPSS and R (R Core Team, 2022) were almost equally popular, mentioned 93 times and 90 times respectively. JASP (JASP Team, 2020), Microsoft Excel and Jamovi (The Jamovi Project, 2021) were the next most popular options, mentioned 20, 17, and 12 times. Many other less popular options were shared; in order of popularity: STATA (8 times), MPLUS (6), Python (4), MATLAB (3), AMOS (3), Process (3), Prism (2),

SAS (2), with LISREL, ESCI, Statistica, VassarStats, JuliaStats and RevMan each mentioned once.

5.4.2 Part 2: True-False Testing

Knowledge varied across the different true-false statements, with no statement achieving more than 64% of correct responses. Table 5.7 shows the breakdown of results for each true-false knowledge statement presented to participants. Note that statement three and statement six are true, and statement one is only considered true according to the flexible definition of a confidence interval (e.g. Cumming, 2012). Statement four, which demonstrates the sample misconception (see Fidler, 2006), scored particularly poorly, with just one quarter of participants correctly identifying it as false. Uncertainty was noticeably high for statement five, which denotes the *confidence-level misconception* (Cumming et al., 2004), which also has an almost equal split of true and false responses. High uncertainty is also apparent for the final statement, with one third of participants responding 'I don't know'. This statement provides the true mathematical definition of a confidence interval.

Table 5.7

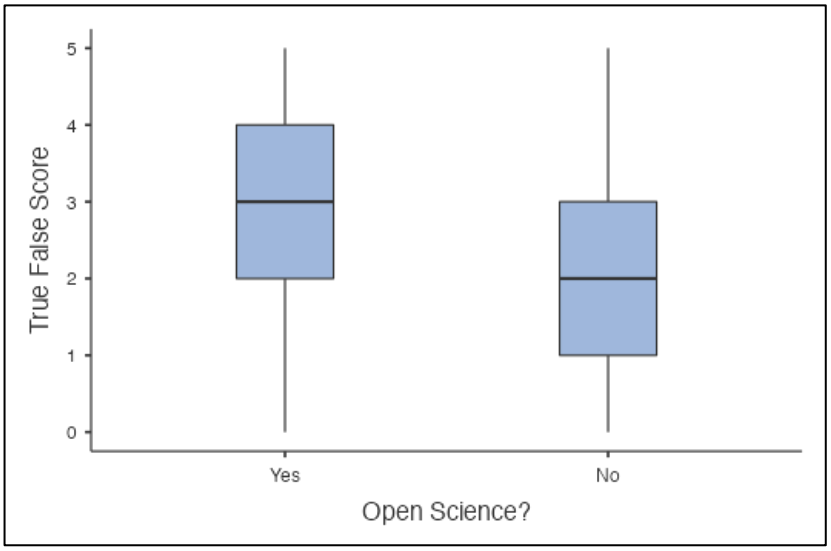
Response Frequencies For Each True-False Knowledge Statement (n = 178-180)

	Statement	***True***	***False***	***I Don't Know***
1	A 95% confidence interval is the range of values for which you are 95% confident that the population mean falls within.	**104** *(58.1%)*	**62** *(34.6%)*	13 *(7.3%)*
2	If all other factors are held constant, an 80% confidence interval will be wider than a 95% confidence interval.	56 *(31.1%)*	**102** *(56.7%)*	22 *(12.2%)*
3	If all other factors are held constant, a confidence interval from a sample of n=25 will be wider than a confidence interval from a sample of n=100.	**114** (63.7%)	35 *(19.6%)*	30 *(16.8%)*
4	A confidence interval gives you the range of plausible values for the true sample mean.	111 *(62.4%)*	**52** *(29.2%)*	15 *(8.4%)*
5	If an experiment is replicated with new samples from the same population, 95% of future means will fall within the original 95% confidence interval.	70 *(39.1%)*	**73** *(40.8%)*	36 *(20.1%)*
6	If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean	**96** *(53.6%)*	23 *(12.9%)*	60 *(33.5%)*

Note. Correct answers are highlighted using **bold underlined** text.

Figure 5.3

Difference in True-False Scores by Open Science Category



The mean number of statements correctly identified as true or false was 2.43 (SD = 1.4) out of 5 (based on statements 2-6), with a median and mode of 2. Participant scores ranged from 0 (18 participants) through to 5 out of 5 (14 participants). The distribution of "I don't know" responses was heavily positively skewed (M = 0.98 out of 6 statements, SD = 1.3), with just nine participants choosing it four or more times. There was no significant difference on true-false scores between job roles, $F(7, 162) = 1.28$, $p = .265$, $n^2 = 0.052$. There was, however, a significant relationship between open science category and score, as shown in the box plot in Figure 5.3. Participants who responded *yes* to the demographic question about open science and psychological gave more correct answers (M = 2.76, SD = 1.40) than their non-open science counterparts (M = 1.98, SD = 1.28), $t(167) = 3.61$, $p < .001$, $d = 0.569$.

When asked, 56% of participants reported not feeling that they have had sufficient training related to the use of confidence intervals, compared to 40% claiming to have had enough training already. However, just 10.5% of the sample expressed no interest in further training, suggesting that there is a widespread appetite for education regardless of prior experiences. Participants typically acknowledged factors such as time and cost as being barriers to engaging with training, and frequently commented on wanting context-specific support that related to their research area. One participant, who responded *maybe* to wanting future training, expressed the following concern:

"I think that it is important, but it feels like the training isn't a necessity until journals want or require the use of confidence intervals"

This highlights the important role that the publishing industry can play in statistical reform.

5.4.3 Part 3: Defining 'Confidence Interval'

The frequencies of correct and incorrect definitions of confidence intervals are shown in Table 5.8, along with frequencies of *I don't know* responses and a breakdown of demographic group differences. Ratings of correctness are shown according to both the strict definition of confidence intervals (e.g. Morey et al. (2016a)) and also when accepting the flexible perspective taken by others (e.g. Miller & Ulrich (2016)) as correct. Very few definitions were scored as correct using the strict criteria (just 31 out of 203 definitions).

Table 5.8

Categorisation of Definitions of a "95% Confidence Interval" (n = 203)

	Strict		***Flexible***		
	Correct	*Incorrect*	*Correct*	*Incorrect*	*'I Don't Know'*
Full Sample	31	146	78	99	26
Job Role					
MSc Student	1	7	3	5	2
Research/Teaching Assistant	1	1	1	1	0
PhD Student	6	48	22	32	8
Postdoctoral Researcher	5	23	10	18	2
Lecturer/Senior Lecturer	11	34	24	21	2
Professor	5	8	10	3	0
Other	0	3	0	3	0
Prefer not to say/Missing	*2*	*22*	*8*	*16*	*1*
Open Science					
Yes	23	72	47	48	7
No	6	51	22	35	8
Prefer not to say/Missing	*2*	*23*	*9*	*16*	*11*

Examples of definitions which were scored as correct under the strict mathematical interpretation of a confidence interval include: "*if we were to repeat this same procedure infinite times, 95% of the time the confidence interval constructed would contain the true population parameter the procedure is estimating*" and "*it is a description of the uncertainty around a statistic in the form of a range of values such that 95% of intervals constructed in a similar way will contain the true parameter value*". In comparison, using the more flexible

scoring approach, correct definitions include: "[a] *95% confidence interval means that one can be 95% confident that the range of values contains the true mean of the population*" and "*the range for which we are 95% "confident" that the true value lies within".*

The use of the phrase *true value* instead of population value was widespread, with 46 participants using this or a similar phrase instead of explicitly mentioning population and sample values. A further 50 participants made no reference of populations, samples, or true values at all (all of which were rated as incorrect), typically providing vague definitions such as "*using this method, the effect will be within this interval in 95% of the cases*".

Incorrect Definitions

Multiple common mistakes were identified in the 99 incorrect definitions (those rated as incorrect by both the strict and flexible standards). Twenty-nine of these definitions included the confidence level misconception (see Cumming et al., 2004), such as "*when running this experiment 100 times, in 95 of these runs, the mean will lie between these 2 numbers*". Twelve of the other incorrect definitions appeared to combine or confuse confidence intervals and NHST, making references to statistical significance ("*the statistically significant value falls within this range*"), the null hypothesis ("*we are 95% confident that we would obtain the results observed here if the null hypothesis is false*") or most frequently, referring to finding results by chance (e.g. "*there is a 5% probability that the result is due to chance*"). In addition, one participant defined a confidence interval as a range of possible *p*-values.

Three further participants mistakenly defined confidence intervals using the sample misconception (see Fidler, 2006), such as "*95% confidence interval refers to the range of values which the sample effect possibly lies within*", while another six participants implied the same mistake using less clear language, such as "*the range where 95% of your results are likely to fall*". The remaining incorrect definitions typically used vague or confusing wording, failed to acknowledge uncertainty, or presented entirely incorrect ideas, such as "*you predict that if the event was run 100 times 95 times your prediction will be right*", "*the bounds that we can be sure the true mean lies within*" and "*having a range of values which may determine that your data is 95% representative of the population*".

5.5 Discussion

Confidence intervals are argued to be a useful statistic which offer a range of plausible values for the population of interest, and highlight the uncertainty of a single sample value. Misconceptions related to confidence intervals have been identified in a variety of settings (e.g. Fidler et al., 2004; Hoekstra et al., 2014; Crooks et al., 2019). The study reported in this chapter makes a novel contribution by examining knowledge and misconceptions using a new true-false knowledge measure, in a contemporary sample of psychological researchers. In addition, this is one of the first studies to collect and report free-text responses to capture the varied knowledge and experiences of psychology researchers working with confidence intervals.

5.5.1 Using, or not Using, Confidence Intervals

Confidence interval use was widespread in this sample, and is significantly higher than reporting rates identified in reviews of the literature (e.g. Fritz et al., 2012). Rates of not using confidence intervals at all reinforce earlier findings by Badenes-Ribera et al. (2018), who also found that 10% of participants reported 'never' using them in their work. However, overall use appears to be higher in the sample measured here, with half of participants reporting that they do use them in all suitable circumstances, as opposed to just 27.7% of participants reporting using them "*quite frequently*" in the work of Badenes-Ribera et al. (2018). It is hard to make direct comparisons as earlier work did not provide an 'always use' response option to participants, but broadly speaking, confidence interval use either remains at similar levels or is higher in this specific sample.

Several common explanations emerged for not, or not always, using them. Particularly notable barriers to confidence interval use were a lack of knowledge about confidence intervals, and the influence of colleagues and supervisors. Participants also acknowledged being strongly influenced by their wider academic peer group, arguing that they don't report confidence intervals because other papers that they have read do not report them. Another common argument was that journal requirements, or a lack of journal requirements, directly influences behaviour, as mentioned by 13 participants. This theme emerged again when participants were asked about interest in training. However, the impact of journal requirements does not appear to be universal: when participants were asked about what

motivated them to include confidence intervals in published papers, just one quarter reported doing so due to journal requirements. Nonetheless, several participants also indirectly connected their behaviour to journals, making comments on limited word counts and space for figures or large tables, suggesting that journal restrictions influence reporting behaviour through multiple means. It is important to consider that this also illustrates the perception of confidence intervals as an add-on to other information, as researchers who use excuses such as word counts do not appear to prioritise reporting wider statistical information in their work.

5.5.2 Strict and Flexible Perspectives

Several explanations for not using confidence intervals highlighted the strict frequentist interpretation of a confidence interval. These participants were critical of a confidence interval having any value at all, given that, considering the strict mathematical interpretation of a confidence interval, one interval appears to be very difficult to interpret. Statement 1 in the true-false measure was designed to investigate these differing perspectives by presenting the flexible '*95% confident*' interpretation of confidence intervals. With more than 1/3 of participants labelling it as false, it appears that awareness of the strict interpretation may be growing. However, when looking specifically at the strict interpretation, which is presented in statement 6 ('*If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean'*), it is clear that there is a broader lack of uncertainty as one third of participants answered "I don't know". It is important to note that statements 1 and 6 are both compatible from the flexible perspective. statement 6 is not 'false' under any perspective: it is simply a mathematical fact about confidence intervals and long-run probability. Therefore, participants marking it as false (12.9%) demonstrate some misunderstanding of the mathematical definition of a confidence interval.

The quantitative data from the true-false statements suggests that awareness of the strict frequentist perspective on confidence intervals is growing, with more than one third of participants labelling statement 1 as false, and more than half of participants labelling statement 6 as true. This is evident in the free-text definitions too, where 31 participants provided a strict frequentist definition of a confidence interval, and a further 11 provided ambiguously worded definitions that could be indicative of a similar understanding of long-run probability. However, it should be acknowledged that one third of participants answered

statement 6 with '*I don't know'*, which suggests that a large proportion of the sample lack an in-depth understanding of a confidence interval. It's also important to note that those participants who marked statement 1 as false do not necessarily have a firm grip on theories of probability. This needs to be explored in a far more nuanced context to draw conclusions about specific knowledge related to probability.

5.5.3 Misconceptions of Confidence Intervals

The true-false knowledge statements also highlighted the prevalence of certain misconceptions in this particular sample of psychological researchers. Almost two thirds of participants incorrectly labelled confidence intervals as plausible values for the sample mean, as opposed to the population mean. Some of these responses could be attributed to misreading the question, but this data aligns strongly with previous findings related to the sample mean misconception shared by Fidler (2006). Participants demonstrated a greater understanding of confidence interval width, both in the context of precision and sample size, although mistakes were still made by 31% of participants for statement 2 (comparing 80% and 95% widths) and 20% of participants for statement 3 (comparing sample sizes of 25 and 100). The sample misconception also appeared in several incorrect definitions of confidence intervals. It appears that the misconceptions noted in the work of both Fidler (2006) and Kalinowski (2010) still persists in this contemporary sample.

The confidence-level misconception, which relates to replication and was coined by Cumming et al. (2004), is also apparent in the data collected for this study. Uncertainty related to replication was apparent in responses to statement 5, with an even division of true and false answers and 20% of participants responding "I don't know". It appears that knowledge regarding replication and confidence intervals has not yet become widespread since the work of Cumming et al. (2004), although arguably, knowledge of replication rates is rather more specific than having a basic understanding of concepts such as confidence interval width and the difference between a population and a sample. The replication misconception was also clear in numerous free-text definitions, indicating that many researchers are still confused over what information is provided by a confidence interval.

5.5.4 Study Limitations

Several limitations must be acknowledged for this research. Firstly, only 10% of participants surveyed in this study reported never using confidence intervals, which is remarkably low in

contrast to reviews of confidence interval reporting across the psychology literature (e.g. Fritz et al., 2013). It is highly likely that self-reported data overestimates behaviour when compared to actual reporting practices. In addition, survey work into the use of any particular statistic would benefit from more nuanced response items, as the data related to confidence interval use (*yes*, *not always*, or *no*) does not provide a rich insight into behaviour.

Furthermore, the sample demographics are somewhat biased towards open science researchers and younger researchers, considering that only 6.3% of participants identified themselves as professors or equivalent tenured staff, and 50% of participants reported engaging in some form of open science or psychological reform. While it is impossible to know how many academics engage in some form of open science or reform behaviours, it is likely that they are over-represented in this sample, as the 'open science movement', for want of a better term to describe psychologists interested in improving the field, has really only spread in the past decade.

The measurement of knowledge related to confidence intervals must also be considered. While this survey took steps to minimise guesswork by including an "I don't know" option, true-false statements present a limited insight into how well confidence intervals are understood, and the statements selected here only represent a few aspects of the topic. Future work could provide longer questionnaires that span various concepts related to confidence intervals, and also examine confidence interval understanding in more applied settings, to look at knowledge and understanding in broader contexts.

5.5.5 Future Directions

The research presented in this chapter indicates that, while confidence interval use may be growing, misconceptions still persist about the basic nature of a confidence interval, such as confusion over whether one interval represents plausible values for a sample or for a population. Future research could adopt several approaches to examine confidence interval knowledge, such as identifying ways to reach more representative samples, and examining knowledge using different measures and different contextual settings.

Beyond research, changes could also be made across both the educational and publishing spheres, to further influence researcher behaviour. While not universal, many participants agreed that they are influenced by journal requirements, and so it is logical to suggest that

more journals should adopt statistical policies which require researchers to expand statistical reporting beyond NHST. The review presented in Chapter 2 indicates that 68 of the top 100 psychology journals request or require authors to include confidence intervals, or similar measures of precision, which may positively impact behaviour. In addition, improved statistical training should be incorporated at all levels of psychology education, with added discussions about the disagreements between researchers who do and do not accept the interpretation that you can be 95% confident in the values of a single confidence interval. Equipping researchers with information like this will enable them to make their own educated decisions about how to interpret confidence intervals.

5.6 Conclusion: The Messy Reality of Confidence Intervals

If confidence intervals are promoted as an improvement on NHST, which has historically been misused and misunderstood, then they must be a clear improvement that is used and understood appropriately by researchers. Therefore, the ideal researcher should not only use them, but also possess basic knowledge about them, including but not limited to the facts shared in the true-false test within this chapter. In addition, an ideal researcher should have a sufficient level of understanding which allows them to appropriately interpret confidence intervals, which requires wider knowledge of the term 'probability' and the mathematical discussions that surround confidence interval interpretation.

Already, the literature indicates that confidence intervals exist within a much less ideal reality, both with regards to individual knowledge, but also at a conceptual level. This conceptual mess is reflected here in the division between researchers who have adopted the strict frequentist perspective versus those who still maintain the '95% confident' approach, and the third more problematic group who still understand a single interval as having a '95% chance' of containing the true population value. This messy reality extends to basic knowledge too, given that participants indicated confusion over the difference between a sample and a population. While high self-reported use of confidence intervals may reflect a growth of 'ideal' reporting behaviour, it is clear that everything beyond this is far from ideal. These findings emphasise the importance of education related to confidence intervals, while also raising the question of whether confidence intervals really represent any improvement to statistical analysis at all (a discussion continued further in Chapter 8).

Chapter 6: Confidence Interval Interpretation and Meta-Analytic Thinking

6.1 Abstract

Background: While reporting confidence intervals provides an explicit indication of uncertainty, it is most useful to interpret them to think critically about a set of findings. However, the literature indicates that confidence intervals are not accurately interpreted by researchers, with confidence intervals often mistakenly connected to NHST. This chapter presents research into confidence interval interpretation, including a partial replication of Coulson et al. (2010), to investigate whether interpretations of confidence intervals have improved over time.

Methods: As part of the confidence interval survey detailed in Chapter 5, participants (n = *206*) were asked to interpret two scenarios about confidence intervals, consisting of one scenario describing a single interval, and a second scenario describing a pair of confidence intervals from two similar studies. Neither scenario included any details related to NHST. Results were examined for interpretations of the confidence interval values, any mentions of NHST-related concepts such as p-values or null hypotheses, and evidence of meta-analytic thinking when presented with two studies together.

Results: For both scenarios, more than half of participants exclusively discussed the confidence intervals in their interpretations. However, a further one third of participants still referred to NHST in their responses. Equal numbers of responses using exclusively-CI or exclusively-NHST approaches indicated a belief that the findings are fixed (e.g. 'there is an effect'). When presented with two studies together, there was an equal division of participants who rated the two studies as 'similar' versus rating them as 'conflicting'.

Conclusion: This contemporary revision of Coulson et al. (2010)'s research presents similar findings a decade later, with participants using a wide variety of approaches to confidence interval interpretation. Despite confidence intervals being proposed as a measure of

uncertainty, which is independent of NHST, many researchers still adopt a fixed mindset when interpreting them, or incorporate NHST into their interpretations.

6.2 Introduction

Chapter 5 of this thesis examined confidence interval knowledge using a novel true-false measure, which briefly studied confidence interval interpretation using simple quantitative questions. However, this data does not offer evidence of how confidence intervals are interpreted within the context of an actual research finding. In addition, quantitative data does not offer a detailed insight into the thought processes employed by researchers when interpreting data. This chapter presents the findings of a qualitative investigation, which examines the conclusions made by researchers when asked to interpret two scenarios that include confidence intervals.

6.2.1 Interpreting Confidence Intervals

Several researchers have conducted research similar to that presented in Chapter 5, studying understanding through the presentation of true-false measures (e.g. Hoekstra et al., 2014; Lyu et al., 2018). However, multiple choice responses provide limited insight into how researchers actually interpret confidence intervals when using them to understand research. One of the first studies to examine interpretation in the context of research results, using qualitative data from participants, was that of Coulson et al. (2010). In their study of academics (which included psychology researchers), they concluded that interpretation of confidence intervals was broadly poor. In particular, researchers often adopted an NHST approach to interpretation even when a set of results did not include *p*-values, by making explicit connections to some element of NHST (e.g. mentioning statistical significance), even though it is an independent concept.

In a similar study just a few years later, Hoekstra et al. (2012) also compared how results are interpreted when presented using NHST versus confidence intervals, using a sample of 66 psychology PhD students. They found that, when shown results as confidence intervals (without any *p*-values), participants were more likely to comment on effect sizes, less likely to mention statistical significance, and less likely to incorrectly accept the null hypothesis (claim that 'no effect exists') in comparison to the NHST formats. However, 41% of participants did still incorporate NHST logic into their interpretations of confidence intervals despite no *p*-values being included, much like the participants in Coulson et al (2010)'s work.

More recently, Crooks et al. (2019) incorporated confidence interval interpretation into their wider conceptual assessment of confidence intervals, which has so far been tested with a sample of 40 undergraduate and postgraduate psychology students. They found that NHST was frequently mentioned by participants when trying to explain how to interpret a confidence interval, similarly to participants in both Coulson et al.'s (2010) and Hoekstra et al. (2012)'s studies. They explored interpretations in more detail to identify other mistakes made, with several misconceptions being shared by participants. The confidence level misconception, which is the mistaken belief that 95% of future replication means will fall within the original study confidence interval (see Cumming & Maillardet, 2006), was the most frequent mistake made. This was followed by the fixed interval misconception, which is the mistaken belief that a confidence interval is a fixed interval, within which a moving population parameter may or may not fall.

The work of Coulson et al. (2010) also examined how researchers interpret a pair of confidence intervals from two studies. One of the two confidence intervals used crossed zero, but, when combined, both intervals together provided evidence in favour of a positive effect. This was intended to examine meta-analytic thinking: the process of combining evidence from multiple sources to generate more reliable conclusions. Meta-analytic thinking is promoted as a particular advantage of the estimation approach, as the use of confidence intervals (which acknowledge the uncertainty of one study) should encourage researchers to look for multiple studies to combine evidence (e.g. Coulson et al., 2010; Cumming, 2012). Coulson et al. (2010) found evidence which lends some support to this: when participants referred to NHST in some way within their interpretation of the two studies (therefore straying back to a traditional, dichotomous mindset), they were highly likely to rate the two studies as having conflicting results. In contrast, when participants focused on the confidence intervals, they were highly likely to rate them as both providing evidence in favour of an effect.

6.2.2 Chapter 6 Overview

The study reported in this chapter extends the research shared in Chapter 5, to examine confidence interval interpretation in a more relevant research context. This study builds on earlier work by Coulson et al. (2010) and Crooks et al. (2019), to examine whether confidence interval interpretation has improved in the past decade in the psychology researcher population.

Objective 1: To examine the approaches that researchers take when interpreting confidence intervals, both to examine whether researchers still incorporate NHST logic (as found in previous research), and also to capture the variety of ideas and information that researchers use to draw conclusions about a set of results.

Objective 2: To examine meta-analytic thinking in a contemporary sample, replicating the earlier work of Coulson et al. (2010). This has two purposes: (1) to see whether researchers naturally combine evidence from two separate studies to draw an overall conclusion and (2) to see whether researchers draw a mathematically accurate conclusion when combining two confidence intervals.

Objective 3: To explore misconceptions and mistakes made by researchers when interpreting confidence intervals, to look for further evidence of the misconceptions that are discussed in Chapter 5.

6.3 Methodology

The data for this project was acquired from the confidence intervals questionnaire reported in Chapter 5. Details of ethical approval, participant recruitment and the wider structure and delivery of the questionnaire are available in Chapter 5, section 5.3.8, and further documentation can be found in Appendix D.

6.3.1 Materials and Procedure

In the final section of the questionnaire reported in Chapter 5, all participants were shown two written scenarios about confidence intervals (see Figure 6.1), and were asked to share their interpretation of the results in free-text boxes, or write "I don't know" if preferred. Scenario 1 described the results from a single study, presenting a mean difference and a single confidence interval, and was written for this study. Scenario 2 described the results from two similar studies, presenting two confidence intervals for the purpose of comparison. This scenario is the two-study written confidence interval format used in Experiment 1 in Coulson et al. (2010), with edited surnames. Note that the surnames in this version are

different because the original Coulson et al. study scenarios used my surname (Collins), which may have confused some participants in this study if I had also used it.

Figure 6.1

Scenarios Presented to Participants Within the Wider Questionnaire

Scenario 1

A study (n=42) reports that the mean weight loss and 95% confidence interval for a longitudinal diet plan is 4.65kg (-1.95, 11.25).

Scenario 2

Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Both Skinner (2018) and Miller (2019) used two independent equal-sized groups and reported the difference between means for the group assigned to the new treatment and the group who maintained their existing treatment.

Skinner (2018) with total n=44 found the new treatment found the difference in means was 3.61 (95% CI: 0.61 to 6.61). The study by Miller (2019) with total n=36 found that the difference in means was 2.23 (95% CI: -1.41 to 5.87). A positive difference indicates a positive outcome for the new treatment.

The original Coulson et al. (2010) study presented participants with several conditions which presented scenarios in different ways, including two graphic presentations, and a written scenario that used NHST instead of confidence intervals. The choice to use the written format of their confidence interval scenario for this study was made because confidence intervals are most commonly presented in written form within published articles, and so this is the most relevant context that can be compared to reading the results of an actual piece of research.

6.3.2 Data Tidying and Analysis

Overall, 172 participants gave a response to Scenario 1, and 155 participants gave a response to Scenario 2. However, for Scenario 1, 16/172 participants wrote '*I don't know'*, and were excluded from further analysis. Similarly, 19/155 participants wrote '*I don't know'* for Scenario 2, and were also excluded. The remaining interpretations of both scenarios were analysed using a basic content analysis, coding each interpretation deductively (top-down) using a codebook (described in more detail below). A more in-depth description of content analysis can be found in Chapter 3, Section 3.3.6 (Drisko and Maschi, 2015). The codebook for both scenarios was developed based on analyses conducted by Coulson et al. (2010) and Hoekstra et al. (2012) in their examination of confidence interval interpretations, and was then expanded based on multiple readthroughs of the data collected for this study.

Coding Scenario 1

First, all responses were coded for interpreting the single scenario using a *confidence interval (CI) approach*, an *NHST approach*, a *CI and NHST approach*, or mentioning *neither.* Examples of all four of these approaches are shown in Table 6.1. This coding strategy differs slightly from Coulson et al. (2010) who only used the categories '*CI-as-NHST'* and '*not-CI-as-NHST'*, in order to provide more insight into the different ways that participants interpreted the scenarios in the study reported here. Note that the purpose of this coding process was to categorise the approach taken to interpreting the scenario, not to score the correctness of the interpretation.

Table 6.1

Example Coding of Scenario Interpretations

Interpretation	Description	Example
CI approach	Only interprets the confidence interval, e.g. explicitly mentions at least 1 of the values	*"The mean for this sample is 4.65,but the true population mean is likely between – 1.95 (ie an effective weight gain) and 11.25 kg."*
NHST approach	Only interprets results using NHST logic, e.g. mentions statistical significance, null hypothesis, or uses zero as a measure of 'no effect'	*"Mean weight loss is not significantly different from 0 at alpha = .05"*
CI and NHST approach	Combines confidence interval and NHST approaches	*"The reported result is 4.65 but could vary between -1.95 and 11.25. It is a very large confidence interval and it includes 0 which indicates it is not significant"*
Neither	Typically vague interpretations which do not clearly discuss the interval or anything related to NHST	*"The mean looks to be a good representation of the possible range of values"*

The analysis approach of Hoekstra et al. (2012) was also integrated into this study, with each interpretation coded for mentions of accepting an imagined null (Fixed – H_0; e.g. 'there is no effect') or an imagined alternative hypothesis (Fixed – H_A; e.g. 'there is an effect' or 'the effect is…'). Note that both of these styles of response neglect the uncertainty of a single set of results, indicating that the reader has not appropriately understood the confidence interval. This was expanded to also record a further 'fixed thinking' approach, using Fixed – Not Sig

to code responses such as "*the CI crosses zero, so the null hypothesis should not be rejected*" which also fails to acknowledge uncertainty without being as explicit as to state that no effect exists.

In line with earlier discussions in this thesis related to the strict and flexible interpretations of a confidence interval, responses were also coded for mentions of the 95% confident approach to interpreting confidence intervals (e.g. Cumming, 2012; defined as the flexible approach in Chapter 5); and for mentions of the long-run probability interpretation (e.g. Morey et al. 2016a; defined as the strict approach in Chapter 5). In addition, each response was coded with *yes, no* or *maybe* to indicate the presence of a mistake or misconception in the interpretation shared. The code *yes* was used for clear mistakes, such as interpreting the confidence interval as a standard deviation. The code *maybe* was used for interpretations that weren't clear enough to be coded as *yes,* such as participants who did not make it clear whether they were talking about plausible population values or sample values. For example, the response "*may change weight anywhere from ~ -2kg to may gain ~ +11kg with 95% certainty*" was coded as *maybe* as it does not clearly identify who may change weight (sample versus population).

After several readthroughs of the data, the codebook was expanded to capture recurring ideas and comments made by participants. A list of phrases that the participant used to describe the "95% ____" was produced (e.g. "*95% likely*"), and responses were coded for mentions of the width of the interval, and how participants interpreted the interval crossing zero, as these were frequently discussed. An illustration of the coding approach used for Scenario 1 is shown in Figure 6.2.

Figure 6.2

Codebook Used to Examine Interpretations of Scenario 1

Approach	H_0 or H_a	Strict or Flex	Mistake?	Mistake Type	Language Used	CI Width	Mentions Zero	Other Comments
CI, NHST, CI-and-NHST, Neither	*H_0, H_a, Fixed – Not Sig, neither*	*Strict, Flex, neither*	*Yes, Maybe, No*	*Free text*	*Probability, Certainty, Likely & Other*	*Wide, No Comment & Other*	*Free text*	*Free text*

Coding Scenario 2

Interpretations of Scenario 2 were coded using a slightly modified version of the same codebook. As Scenario 2 presented the results of two studies together, responses were coded for mentions of the two studies (and their accompanying confidence intervals) as being similar, or different, which is the same approach used to analyse the same scenario by Coulson et al. (2010). For the study analysed here, a third code choice of *no comment* was added to label participants who did not connect the two studies together. After an initial readthrough of the data, the **'**CI Width**'** variable was discarded as it was not relevant to responses, and the "mentions zero" variable was discarded as it did not provide richer data than was already captured using the other variables. The codebook used to analyse Scenario 2 is shown in Figure 6.3.

Figure 6.3

Codebook Used to Examine Interpretations of Scenario 2

Approach	Compatible?	H_0 or H_a	Strict or Flex	Mistake?	Mistake Type	Language Used	Other Comments
CI, NHST, CI-and-NHST, neither	*Similar, Different or no comment*	*H_0, H_a, Fixed – Not Sig, neither*	*Strict, Flex, neither*	*Yes, Maybe, No*		*Probability, Certainty, Likely & others*	

6.3.3 Inter Rater Reliability

A second coder, RW, independently coded a randomised subset of 20% of the responses to Scenarios 1 and 2. For Scenario 1, Cohen's kappa was initially 0.709, which was revised upwards to 0.819 after discussions to clarify that mentioning zero was not always indicative of an NHST approach. For Scenario 2, Cohen's kappa of interpretation approach judgements was initially 0.891, which was revised upwards to 1.0 (perfect agreement) after identifying two missed instances of NHST comments by EC. The Cohen's kappa score for judgements of the scenarios being compatible or not was 0.947. When judging kappa, scores of .61 – .80 correspond to 'substantial agreement' and scores >.80 correspond to 'almost perfect agreement' (Landis & Koch, 1977). Note that for any remaining disagreements, EC's decision was determined to be the final judgement, as the primary researcher.

6.3.4 Participants

Of the 206 participants in the overall confidence interval questionnaire study (Chapter 5), 172 interpreted one or both scenarios, and so are included in the data set for this study. Of these 172 participants, 103 identified as engaging with open science in some way, compared to 66 who did not (with 3 missing responses). Participants were from the same assortment of job positions, psychology sub-fields and countries as reported in Chapter 5, section 5.3.8.

6.4 Results

6.4.1 Scenario 1: A Single Interval

Of the 156 interpretations of Scenario 1, 86 were categorised as interpreting using a confidence interval (CI) approach, 20 as an NHST approach, 29 as CI-and-NHST together, and 21 as neither.

The majority of interpretations presented some indication of uncertainty, either by using language which indicated that the results are simply plausible findings, such as "*the mean for this sample is 4.65,but the true population mean is likely between - 1.95 (ie an effective weight gain) and 11.25 kg*", or by referencing uncertainty directly: "*the CI crosses zero so, if we assume that the CI contains the population mean, it could feasibly be 0. There is quite a lot of uncertainty in the estimate*". However, 37 participants indicated a perception that the results are definite, either by mistakenly declaring the presence of an effect (coded as accepting H_a) such as "*the true estimate is between those two values*", or conversely interpreting the findings as demonstrating no effect (accepting H_0,); for instance, "*the confidence interval includes zero, which means no true effect*".

Nine of these 37 participants did not go as far as to declare that no effect exists, but did exclusively interpret the results as being statistically non-significant, without offering any further comments that would suggest a consideration of uncertainty or a range of possible values (e.g. "*there was no significant effect of diet plan on weight loss (as confidence intervals include the null result - i.e. zero weight loss)"*). The frequency of fixed interpretations per approach is shown in Table 6.2, which shows that both the CI and NHST approaches lead to equal numbers of interpretations which perceived the results as certain.

Table 6.2
Fixed Interpretations Within Each Overall Approach

	CI	NHST	CI-and-NHST	Neither
Fixed - Accepts H_a	13	0	4	0
Fixed - Accepts H_0	2	6	1	2
Fixed – Not Significant	0	9	0	0

Neither the strict long-run probability interpretation or the flexible '95% confident' interpretation appeared frequently within participant responses, with the former mentioned or implied just 4 times, and the latter mentioned by 8 participants. While '95% confident' was the most popular phrase used, participants also opted for '95% probability' (*n = 5*), '95% chance' (*n = 4*), '95% certain' (*n = 5*), '95% likely' (*n = 4*) and '95% sure' (*n = 1*).

The interval width and the inclusion of zero within the interval both drew frequent comments from participants. The width of the interval was mentioned in 45 responses, and while most participants simply noted that the range of plausible values was wide, one participant mistook a wide interval for equalling a large effect, and another inversely identified the width as 'not wide'. Several participants connected the width of the interval to increased uncertainty in the results, with one participant making the strong claim that "*the wider CIs generally the less trustworthy the data*". Discussions of the confidence interval spanning zero were more varied, with 33 participants explicitly using zero to interpret the result as not being statistically significant, compared to 17 participants who described zero being just one plausible population value of many. Other participants commented on zero reducing the reliability of the results, using language such as "*no confident claim can be reliably made*" and "*there is not compelling evidence*". One participant made the particularly strong claim that "[the] *confidence interval* [is] *not significant as it crosses zero, so result is meaningless*".

A variety of mistakes and misconceptions were identified in interpretations of Scenario 1, which are presented in Table 6.3, divided into definite mistakes and possible mistakes (where a participant did not provide enough detail to establish whether they were incorrect or not). The most frequent mistake made was defining a confidence interval as either the full range of sample values, or as 95% of the range of sample values. The confidence level misconception, which is the incorrect belief that one confidence interval includes 95% of future means from replication studies (see Cumming and Maillardet, 2006) was also commonly shared by participants. Participants also made several mistakes classed together under 'other', including

getting the relationship between 0 and statistical significance wrong (as quoted in Table 6.3), incorrectly reading the scenario and describing large weight gains, describing the findings as highly positively skewed, and describing (without explanation) a new set of values where future weight loss would fall.

Table 6.3
Confidence Interval Interpretation Mistakes for Scenario 1

Mistake	**Example**	***Frequency***	
		Definite	**Possible**
Fixed Sample Mean	*"With 95% certainty, one loses a mean of 4.65kg using this diet plan"*	4	-
Sample Range	*"On average a participant lost 4.65kg, but this ranged from participants gaining weight to losing up to 11.25 kg"* **-** *"The mean weight loss is very likely between -1.95 and 11.25 kg"*	17	9
Confidence Interval as Standard Deviation	*"This shows that the score is nearly 2 standard deviations below the mean"*	2	-
5% Due to Chance	*"weigh loss would vary between participants putting on 1.95kg and losing 11.25kg on average with a 95% level of confidence (i.e., that 5% of the results would be due to chance and not explained by the intervention"*	2	-
Confidence Level Misconception	*"If repeated samples were drawn from the population, 95% of these samples would report an avg weight loss between -1.95kg and 11.25kg"*	8	-
Other	*"The CI is quite wide, demonstrating a large effect. It also spans zero, which is evidence of a significant result."*	5	2

6.4.2 Scenario 2: Two Intervals

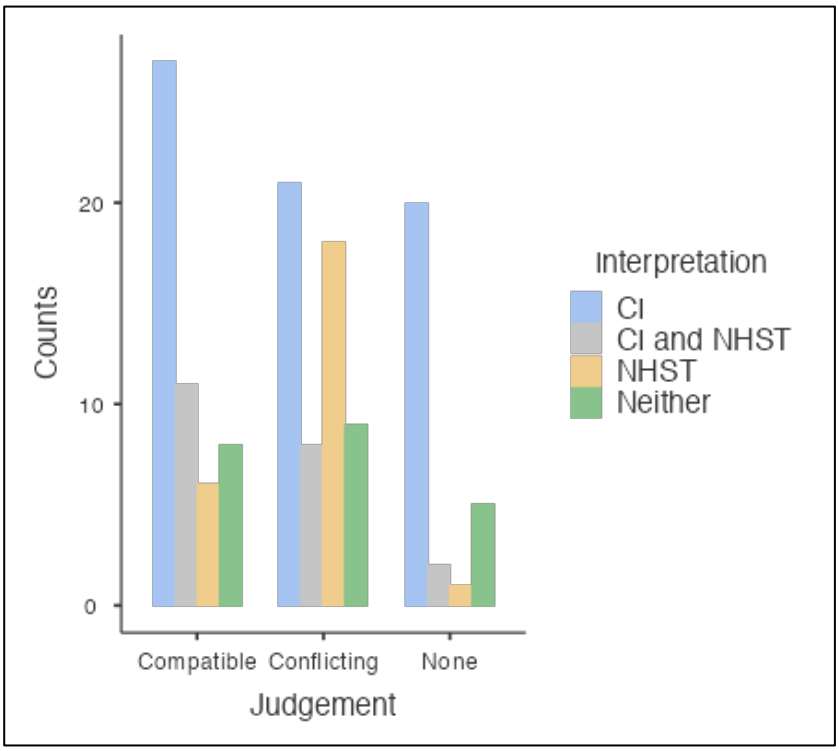
Of the 136 interpretations of Scenario 2, perspectives were almost equally divided. Fifty-three participants described the two results as similar, although just three explicitly commented on using a meta-analytic process to draw any conclusions. In comparison, 54 participants described the two results as conflicting. The remaining 29 participants did not

comment on how the two studies compared to one another. Examples of interpretations categorised as *similar* include "*both studies showed a similar mean difference; a treatment effect favouring the intervention group*" and "*taking the two studies together, the evidence indicates that the new treatment performs better that the existing treatment*". In contrast, *conflicting* interpretations typically referenced the second confidence interval crossing zero as justification for the two results to not be compatible, such as "*Skinner's study finds a positive effect of the treatment on insomnia, while Miller' finds no effect as it crosses the 0 and has a negative CI*".

Sixty-eight responses were categorised as interpreting the results using a confidence interval (CI) approach, 25 as interpreting using NHST, 21 as CI-and-NHST, and 22 as neither. Typical responses categorised as 'neither' were very vague, consisting of interpretations such as "*there is potentially conflicting evidence regarding the new treatment of insomnia more research with larger samples are needed*".

Figure 6.4

Participant Judgements and Interpretation Approaches for Scenario 2



Overall, participants who used an NHST approach to interpreting data were more likely to rate the two sets of results shared in Scenario 2 as conflicting, as shown in Figure 6.4. A chi

square test of independence found a significant relationship between interpretation approach and judgement of the two study intervals, $X^2(6, 136) = 17.3$, $p = .008$, $V = .25$. It should be noted that this is not a particularly large effect size, and overall group differences are quite small, particularly for the participants who used an exclusively-CI approach to interpreting the intervals. Noticeably, 28 of these participants rated the two results as similar, while 21 of them rated them as conflicting, and the remaining 21 made no comment.

When interpreting Scenario 2, zero participants referenced the long-run definition of a confidence interval. Just two participant explicitly used '95% confident' to interpret the intervals, although participants did still use terms such as '95% certain' (*n = 4*), '95% chance' (*n = 2*), '95% likely' (*n = 2*), and '95% sure' (*n = 1*). Mistakes appeared much less frequently than in responses to Scenario 1, although three participants did incorrectly interpret the confidence interval as a measure of the sample values, such as "*there is a 95% chance that the first finding fell between 0.61 and 6.61, and the second between -1.41 and 5.87*". In addition, one participant mistakenly interpreted the interval using logic similar to that of *p*-value interpretations ("*both studies had a confidence interval which I think maybe means that there is a 5% chance that these results are due to chance*"), and two further participants gave vague responses which indicated that they had misunderstood the scenario.

Contrastingly, several participants indicated reflecting more widely on the study to aid their interpretation of the results. Overall, 29 participants commented on the small sample sizes of the two studies, or a potential lack of study power, while several other participants asked for additional contextual information about the study design.

6.5 Discussion

The study presented in this chapter provides a contemporary insight into how well psychological researchers understand confidence intervals (CI). Instead of comparing results presented using CI and NHST formats like earlier work by Coulson et al. (2010) and Hoekstra et al. (2012), this study instead examined how confidence interval results are interpreted in the absence of NHST. In addition, this study incorporated two types of research scenario: one presenting a single confidence interval, and another presenting two intervals from two very similar studies together. The findings presented here reinforce the earlier work

of Coulson et al. (2010), demonstrating that more than a decade later, psychological researchers still interpret confidence intervals in a huge variety of ways, ranging from sensible and insightful evaluations to rigid and unjustified dismissals of possible effects.

6.5.1 Varied Interpretations of Confidence Intervals

When interpreting both scenarios in this study, the 'confidence interval' mindset was adopted most frequently by participants, used exclusively by over half of the sample for both scenarios. Even when participants made a reference to NHST in their responses, more than half of the time they also incorporated a discussion of the confidence interval values. However, equally it must be noted that neither Scenario 1 or 2 made any reference to NHST, and yet almost one third of responses to each scenario incorporated NHST into their response. In Hoekstra et al. (2012, p. 1049), they suggest that "*CIs are not used as an implicit form of significance testing*", but arguably this is not true of the participants in this study, particularly given that a sizeable proportion of participants interpreted confidence intervals exclusively using NHST. These findings demonstrate that NHST still remains ingrained in researchers; much the same as it did more than a decade ago as reported by Coulson et al. (2010).

A particularly striking finding here is that the 'fixed mindset' of firmly (and mistakenly) accepting H_0 or H_a was not limited to those participants using an NHST approach to interpretation: a lack of uncertainty was displayed in equal numbers of exclusively-CI and exclusively-NHST responses. This indicates that, even when interpreted without relying on a dichotomous NHST mindset, confidence intervals do not guarantee improved inferences. This is also reflected in the findings from Scenario 2: despite participants who rated the two results as *similar* being much more likely to use an exclusively-CI approach than any other, many participants using the exclusively-CI approach still described the two studies as conflicting, or made no comment connecting the two studies at all (as shown in Figure 6.4).

6.5.2 Meta-Analytic Thinking

Just one fifth of participants failed to reflect on both confidence intervals together when interpreting Scenario 2, with the remaining four fifths adopting a meta-analytic approach to consider the two results together. Thinking meta-analytically is arguably a strength of a researcher, who looks beyond single pieces of evidence and can begin to put them together to generate stronger evidence bases. However, opinions were almost evenly divided between

those who rated the two scenarios as showing overall similar findings (*n = 53*), versus those who interpreted them as conflicting (*n = 54*). The two intervals shared in Scenario 2, if combined meta-analytically, result in the interval shown in Figure 6.5. It can be seen that in this case, the meta-analytic confidence interval indicates an overall positive effect, conditional upon the two samples being drawn from the same population; and so, the correct interpretation is one which concludes that both studies contribute evidence for an effect (i.e. do not conflict). The mixed findings here suggest that, while the majority of researchers are able to adopt a meta-analytic mindset, they are not yet equipped with sufficient knowledge to do so in a mathematically accurate way.

Figure 6.5

Meta-Analytic Combination of Results From Scenario 2

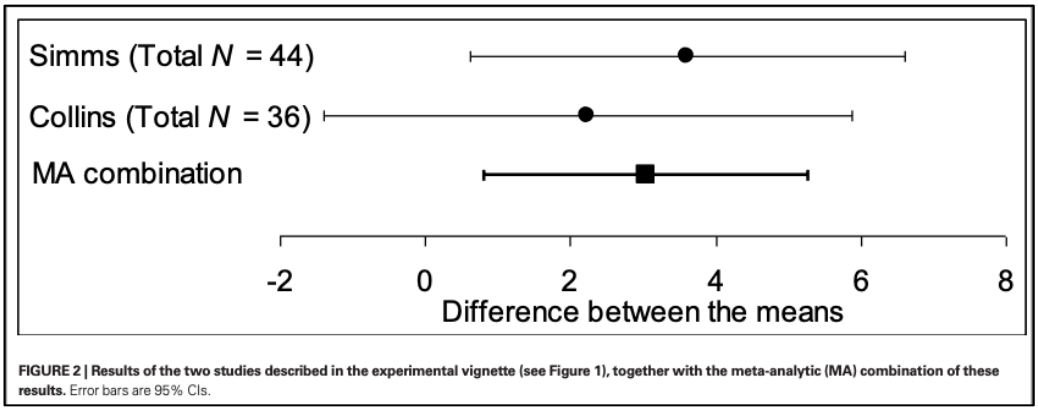


FIGURE 2 | Results of the two studies described in the experimental vignette (see Figure 1), together with the meta-analytic (MA) combination of these results. Error bars are 95% CIs.

Note. From *Confidence Intervals Permit, But Do Not Guarantee, Better Inference Than Statistical Significance Testing*, by Coulson et al., 2010, p. 4. © Coulson, Healey, Fidler & Cumming. Unrestricted sharing with credit permitted.

6.5.3 Interpretations, Misconceptions and Mistakes

When counting mistakes, errors of language (e.g. incorrectly using 'certain' or 'probability') were disregarded in favour of identifying mistakes that indicate a wider lack of knowledge about confidence intervals. Most frequently, participants misunderstood a confidence interval as being a range of *sample* values, demonstrating that they have not grasped the inferential nature of a confidence interval. Others appear to have combined the concepts of confidence intervals and NHST to create a new hybrid misunderstanding, getting confused over percentages and their relationship with results being attributed to 'chance'.

Thinking more broadly about understanding confidence intervals also raises the question of whether conceptual definitions of a confidence interval relate to how they are interpreted. In Chapter 5, data collected from the same sample revealed 31 instances of confidence intervals

being defined under the strict long-run probability approach, and 47 being defined using the '95% confident' approach. However, in this study, fewer than 5% of responses referred to either of these interpretations, which suggests that there is a disconnect between conceptual knowledge and putting knowledge into practice in the context of research studies.

6.5.4 Study Limitations and Future Directions

Several limitations must be acknowledged when considering the findings of this study. The demographic limitations discussed in previous chapters are still valid here, primarily with regards to the generalisability of these results. It is not straightforward to establish current engagement with various aspects of open science and reform, but it is likely that researchers who have adopted more progressive behaviours, and are more likely to have been exposed to conversations about statistical reform, are over-represented in this research. The sample is also biased towards early career researchers, particularly PhD students and postdoctoral researchers, and so does not provide as much insight into the statistical abilities of those who have more career experience in research. However, if confidence interval interpretation is diverse and often inaccurate in this sample, it would not be unreasonable to suggest that it is even less accurate in samples who do not engage with statistical reform and likely stick to more traditional data analysis. It is also unknown what more data from those with more experience would uncover: there could be an even deeper level of ingrained NHST-thinking; or there could be more understanding simply due to a longer career of being exposed to varying statistical concepts. Studying how those with a longer career interact with confidence intervals, and indeed other statistics such as effect sizes, may offer additional insight into the influence that prolonged exposure to NHST has on statistical practices and interpretations.

The method used should also be evaluated. Providing two short scenarios and two free-text response boxes does not necessarily translate to how researchers may interact with published articles, particularly when the topics or methods in Scenario 1 and 2 may not have any relevance to their own work. Indeed, interpretations may be improved with more familiar contexts, which would be a valuable future direction for research to follow. In addition, online questionnaires do not facilitate nuance and do not allow for easy follow up questions and discussions which may highlight details and understanding that has been missed here. It should also be noted that data was collected during the COVID-19 pandemic, which has been a period of stress and uncertainty for most, and is highly likely to have negatively influenced how much engagement and energy participants gave to this research.

While future research will offer more informative insights into researcher abilities, the findings reported here reinforce the importance of education related to confidence intervals. Given that interpretations were hugely varied and often unique, it is obvious that researchers would benefit from improved resources and support if they are to appropriately use confidence intervals in their work. However, this chapter also reinforces the conclusions drawn in Chapter 5, further highlighting that the most appropriate future direction for confidence intervals may instead be not to use them at all, given that they do not appear to be any better used or understood than NHST.

6.6 Conclusion: The Messy Reality of Confidence Intervals (Part II)

As discussed in Chapter 5, the ideal researcher should not only use confidence intervals, but also have both basic knowledge and a good understanding of them. Confidence intervals should not just be reported, but also interpreted, with an ideal researcher offering sensible interpretations. These ideal interpretations should offer an accurate reflection of the interval values, avoid incorporating NHST (e.g. by not using a confidence interval as a measure of statistical significance) and also acknowledge that the presence of zero does not provide concrete evidence that no effect exists. More implicitly, the ideal researcher should be able to interpret a confidence interval using their preferred approach to probability, based on an awareness that there are differences of perspective (as discussed in Chapter 5) with regards to being '95% confident'. However, in contrast to this ideal scenario, the findings of this chapter instead reinforce the messy reality of confidence intervals identified in Chapter 5. Interpretations were hugely varied and often incorrect, and often neglected the confidence intervals entirely by adopting an NHST mindset and only sharing comments about statistical significance and null hypotheses.

Perhaps more concerningly, the data shared in this chapter highlights another important conceptual 'mess' regarding uncertainty versus fixed mindsets. Confidence intervals are promoted as part of the 'estimation approach', where the ideal researcher would use them to critically evaluate and reflect on the uncertainty of a single finding. However, where participants in this study demonstrated fixed thinking (e.g. with statements such as 'no effect exists'), they were equally as likely to do so using a confidence interval approach as those

who used an NHST approach, suggesting that confidence intervals do not naturally invoke reflections on uncertainty. It is apparent based on both this data, and the findings of Chapter 5, that researchers do not have a confident understanding of confidence intervals.

Part 2: Power and Power Analysis

Chapter 7: Using and Misunderstanding Power in Psychological Research

Sections of this chapter also feature in Collins, E., & Watt, R. (2021). Using and Understanding Power in Psychological Research: A Survey Study. Collabra: Psychology, 7(1), 28250. https://doi.org/10.1525/collabra.28250

7.1 Abstract

Background: As the use of NHST persists within psychology, it is sensible to reduce the risk of Type II errors by increasing the statistical power of studies. As many organisations now encourage researchers to plan their sample sizes using power analyses, it is important to identify any difficulties that researchers experience when trying to evaluate the power of their work. This chapter presents new research into the use of power analyses and knowledge of statistical power in psychology.

Methods: An international sample of psychology researchers (n = 214) completed an online questionnaire about statistical power and power analysis. Participants were asked about their sample size planning approaches, effect size estimation methods, power analysis software preferences and using post hoc power analyses. Knowledge of power was studied by asking participants to define the term 'statistical power' in their own words.

Results: Power analysis use was high, with just 30 participants in this study having never used an a priori power analysis. Participants reported difficulties with computing a priori power analyses for complex research designs, and approaches to effect size estimation. The misconception that a post hoc power analysis computes the actual power of a study was also held by approximately one quarter of the sample. Participants typically could not accurately define power, with 57 participants providing an incorrect definition and just 17 providing an entirely accurate definition.

Conclusions: Despite a priori power analysis use increasing, participants reported several difficulties and barriers in this study which suggest that power analyses may not be used

appropriately. In addition, participants appear to lack a firm understanding of what power is as a concept. Researchers would benefit from clear tutorials and improved educational materials to ensure that power analyses do not become a new tick-box exercise.

7.2 Introduction

Thus far in this thesis, the estimation approach has been explored as a plausible addition or alternative to null hypothesis significance testing (NHST). However, in Chapter 1, statistical power was also introduced as an important concept within statistical reform, as it offers a way to minimise the risk of Type II errors. Given that psychology is unlikely to make dramatic shifts away from NHST as a key method of data analysis, it would be sensible to improve the reliability of findings tested using this framework. Furthermore, Chapter 2 demonstrated that more than one third of the top 100 psychology journals either require or encourage researchers to use an a priori power analysis, or at least consider study power in some form, and so researchers need to be equipped with sufficient knowledge and skill to do so appropriately. This chapter presents the findings of a third questionnaire study, which has examined the use of power analyses and knowledge of power in a psychology researcher sample.

7.2.1. What is Power?

Statistical power ('power') is the probability of obtaining a statistically significant outcome for a test, given a particular alpha level, sample size and population effect size. It is tied to the Type II (false negative) error rate, where error rates increase as power decreases, and the conventionally 'acceptable' level of power in psychological research is 80%, equal to a 20% chance of a Type II error. Reviews suggest that psychological research is consistently underpowered, primarily due to insufficient sample sizes. For example, Cohen calculated the average power of research in one journal to be 18%, 48% and 83% for small, medium and large effects, respectively (Cohen, 1962). As the Type II error rate is equal to $1 - power$, then Type II error rates in the literature reviewed by Cohen could be as high as 82% ($100\% - 18\%$) for research studying small effects. This is particularly important because the effects studied in psychology are often small, based on Cohen's original benchmark of d = 0.2 (De Boeck and Jeon, 2018). Low power means that these small effects may be missed entirely by researchers.

Despite Cohen's longstanding efforts to encourage more researchers to think carefully about the power of their work, contemporary reviews show very little improvement. For instance, Stanley et al.'s (2018) review of 200 meta-analyses calculated a median overall statistical

power of only 36%. However, as more journals implement requirements and guidelines related to statistical power, it is hoped that the overall power of psychological research will improve.

7.2.2 A Brief Overview of Power Analyses

In Chapter 1, the term power analysis was first introduced. This term is typically used to refer more specifically to an 'a priori' power analysis, which is the recommended calculation used for planning research sample sizes with adequate power. This calculation uses a set alpha, intended power level, and estimated population effect size to identify the minimum sample size for a well-powered study. However, other power calculations also exist: sensitivity analyses, and post hoc power analyses. Sensitivity power analyses are often used retrospectively to identify the smallest sample effect size that a particular study could detect, given a fixed sample size, chosen alpha and chosen power level. This process is sometimes favoured as it does not rely on inaccurate estimates of population effect sizes, although it must be noted that it still does not provide any information on the actual population effect size.

Post hoc (or 'observed') power is perhaps the most controversial approach to power analysis. Post hoc power is calculated after data has been analysed, and uses the sample effect size, chosen alpha, and actual sample size to calculate 'study power'. Traditionally, it has been used for two purposes: to claim that a null effect is only due to low study power; or to attempt to rule out an alternate hypothesis, because 'post hoc power' appears to be high. However, as briefly discussed in Chapter 1, this is a problematic calculation. The mathematical relationship between post hoc power and *p*-values means that null results ($p > .05$) will always correspond to power being less than 50% (Yuan & Maxwell, 2005; Lakens, 2014), which can be demonstrated using Fisher's z transformation (Collins & Watt, 2021):

$$pw_{post-hoc} = 1 - normcdf\left(norminv\left(1-\frac{\alpha}{2}\right) - norminv(1-\frac{p_0}{2})\right)$$

In addition, the measured sample effect size is highly unlikely to be a reliable reflection of the true population effect size due to sampling error (Gelman, 2019). Despite its flaws, post hoc power was mistakenly encouraged by many academics at the start of the 21st century (e.g. Onwuegbuzie & Leech, 2004), which may still influence the behaviour and knowledge of

academics today. The review reported in Chapter 2 identified only two instances of top psychology journals explicitly banning the use of post hoc power as of October 2021.

7.2.3 Researchers and Power

While many broad reviews of the literature exist, little research has examined the use and understanding of power from the individual perspectives of researchers. Despite Cohen's early work encouraging the adoption of power analyses (Cohen, 1988), a survey in the late 20th century found that only 36.1% of surveyed psychology and management academics used a priori power analyses in any of their research (Mone et al., 1996). However, 50% of their sample did report employing post hoc power analyses to investigate results that were not statistically significant. Another, more recent, survey of psychologists found that only 47% reported using an a priori power analysis for sample size planning; just a 10% increase since Mone et al.'s (1996) investigation two decades prior (Bakker et al., 2016). When evaluating actual behaviour instead of self-reported behaviour, power analysis use appears to remain much lower: for example, Tressoldi and Giofrè (2015) found that only 2.9% of 853 psychology articles reported an a priori power analysis or discussed sample size. In addition, a review of reported power analyses revealed that they often lack detail about effect size estimation and which software was used to calculate power. Almost half of power analyses failed to justify their choice of effect size, and those that do justify it often rely on estimates from previous literature, or Cohen's general benchmarks (Bakker et al., 2020).

Mone et al. (1996) also briefly examined barriers to power analysis use, with researchers reporting difficulties with software and an overall lack of knowledge about power. The research of Bakker et al. (2016) also suggests that an insufficient understanding of power is a barrier to power analysis use. In a brief knowledge test, three quarters of their sample could identify the correct definition of power when presented with a list of options. However, further testing found that most participants overestimated the power of studies investigating small effect sizes, and underestimated the sample sizes which would correspond to adequate power to detect typical effects in psychology, suggesting that psychologists have incorrect intuitions about power.

7.2.4 Chapter 7 Overview

The intention of this study was to contribute new data on the use and knowledge of power analyses and statistical power in the psychology researcher population. This data will provide

an insight into current behaviours compared to recommendations, and may identify issues that can be addressed to ensure that power analyses are being used correctly within psychology. The specific objectives were as follows:

Objective 1: To explore power analysis use, including possible barriers to using power analyses, approaches to effect size estimation, and use of post hoc power analyses.

Objective 2: To examine knowledge of statistical power, both through the free-text data that participants use to explain their power analysis experiences, and by asking participants to define the term 'statistical power' in their own words.

7.3 Methodology

7.3.1 Ethics

This study received ethical approval from the University of Stirling's General University Ethics Panel (GUEP #864 19-20) and adhered to the Code of Human Research Ethics guidelines of the BPS (BPS, 2014). Documentation can be found in Appendix E.

7.3.2 Sampling and Inclusion Criteria

An a priori power analysis was not deemed suitable due to the exploratory nature of this study. In the same manner as the other questionnaire studies that form this thesis, participants were recruited via opportunity sampling, with the intention of capturing as many responses as possible during a month-long sampling window.

Data was collected between 16th April and 16th May 2020, during the first wave of the COVID-19 pandemic. Participants were recruited through Twitter and within a LinkedIn research group for psychology researchers, and the study questionnaire was shared through academic mailing lists such as the PsyPAG mailing list for UK postgraduates and postdoctoral researchers, the University of Stirling Psychology Staff and PhD mailing lists, and the UK Research Methods Psychology JISC mailing list. Advertisements used for this study are similar to those used in the two previous questionnaire studies, examples of which are found in Appendix B.

As with the previous two questionnaire studies in this thesis, all self-identifying psychology researchers, actively involved in some level of quantitative research, were eligible to take part in this study. Once again, undergraduate students were not eligible to take part in this study as they are not typically responsible for generating published research. Participation was open to researchers in all countries to capture a range of diverse experiences.

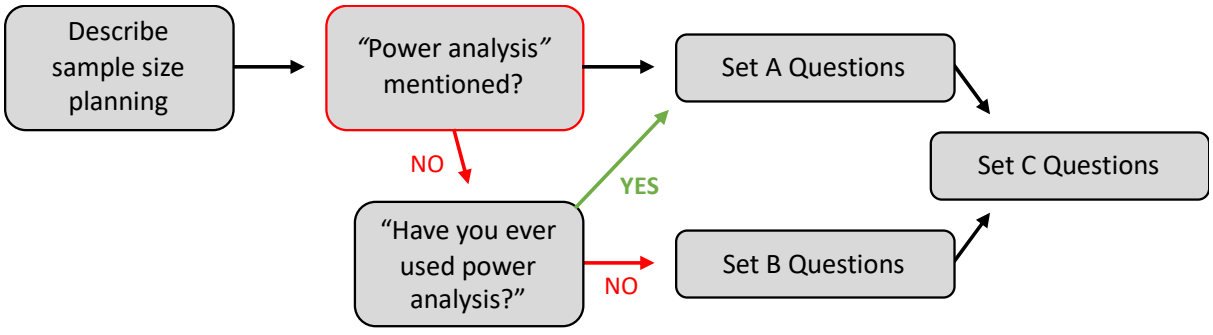
7.3.3 Materials

Data for this study was collected via an online questionnaire (shared in Appendix E). All questions were developed by the researcher, and are detailed below. The questionnaire was first piloted with a professor and two PhD students to ensure that question wording was clear, and no changes were recommended by pilot participants.

Sample Size Planning. Participants were asked "*as a researcher, how do you generally plan your sample size for quantitative research?"* with a large free-text response box. This question was designed to encourage an honest overview of sample size planning with space to report any number of approaches. Qualtrics was set up to identify mentions of the key term *power analysis*. Participants who mentioned the term *power analysis* were automatically routed to the Set A Questions, as shown in Figure 7.1.

Figure 7.1

Order of Materials Presented to Participants



Use of Power Analysis. This question was displayed to participants who had not explicitly used the term *power analysis* in their previous response to sample size planning. These participants were explicitly asked if they had ever used an a priori power analysis for sample size planning, with options *yes* or *no*. This question was used to direct participants to the set of questions relevant to their experiences (Set A or Set B, shown in Figure 7.1).

7.3.3.1 Set A Questions (Experience of Power Analysis)

As shown in Figure 7.1, set A indicates the questions displayed to participants who either reported using analyses in their free-text responses, or to participants who explicitly confirmed having used an a priori power analysis before.

Frequency of Power Analysis. Participants were asked to estimate how often they used power analysis in circumstances where it would be suitable, worded as "*how frequently do you use a priori power analyses, as a percentage of the studies you conduct that include hypothesis testing (p-values)?"*. Responses were measured on a 0-100 sliding scale. If participants did not choose '100', the follow-up question "*why do you not use power analysis 100% of the time for suitable studies?"* was displayed. This question was used to quantify whether power analysis is a process that is typically employed for all relevant study planning; and if not always used, why not.

Software Use. Participants were asked to list the software that they use to calculate power analyses, with space for multiple responses. This question was designed to capture the diverse options that researchers may be using.

Table 7.1
Seven Options Presented to Participants for Effect Size Estimation

	Strategy for Determining an Effect Size
1	Use an effect size from the results of other published literature
2	Use the same effect size as a previous similar study reported in their methods
3	Use a small or medium effect size e.g. Cohen's recommendations
4	Take recommendations from other researchers
5	Use the smallest effect size of interest for my field or "meaningful" effect size for my field
6	Run a pilot study to calculate an effect size first
7	Other

Effect Size Estimation. Participants were shown a list of six different approaches to effect size estimation for a priori power analyses (see Table 7.1), and were asked to select all of the methods they use in their own calculations. A seventh 'Other' option was also available, and participants who selected this were asked to provide more details. This data provides some

insight into whether researchers are making educated choices about effect sizes for power analyses.

Defining Power. Participants were asked to define power in their own words, or write *I don't know* in the free-text box instead. Earlier research has used multiple-choice options to test knowledge of power (Bakker et al., 2016), which provides a limited measure of the ability to differentiate between list items. In contrast, the study presented here asks researchers to draw exclusively on their own knowledge to provide a response. This question was designed to establish whether participants have a firm understanding of what statistical power is.

Post Hoc Power. Participants were asked if they had ever calculated post hoc (or observed) power, and if *yes*, why they had done so. This question was included because post hoc power analysis is mistakenly perceived as a measure of actual study power, and it is not yet known whether or not researchers are aware that its use is now discouraged as it is simply a function of a study's *p*-value.

Importance of Power. Participants were asked to rate the importance of statistical power in psychology research using a four-item Likert scale, with options ranging from *not important at all* through to *very important*. A fifth *I don't know* response choice was available. This data offers some insight into what researchers think about statistical power, as it grows in popularity across the discipline.

7.3.3.2 Set B Questions (No Experience of Power Analysis)

As shown in Figure 7.1, participants who responded *no* to having experience of power analysis, were routed into question set B. They were asked to explain why they haven't used power analyses in their research, in case any particular barriers exist that may be easily addressed by organisations to support researchers. Participants were then shown the following questions from set A: *post hoc power*, *defining power*, and *importance of power*, as detailed in section 7.3.3.1.

7.3.3.3 Set C Questions (All Participants)

All participants were shown the same demographic questions at the end of the questionnaire.

Demographics. The demographic questions asked within this questionnaire match those detailed in the previous questionnaire studies of this thesis (Chapter 3 and Chapter 5). Participants provided information about their job position and country of work, sub-field of psychology, and engagement with psychological reform and/or open science, as measured using the question "*are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar?"*.

7.3.4 Procedure

The online questionnaire for this study was delivered via Qualtrics (Qualtrics, Provo, UT). It began with an information sheet and digital consent form, which signposted that participation was anonymous, and that all questionnaire questions were optional and could be left blank if preferred. All participants provided informed consent before data collection via an electronic check box. Participants were informed that they were free to withdraw from the study at any point with no penalty by closing the browser, and that they had the right to withdraw their data within 14 days of participation by using the unique ID code assigned to them on the information sheet. The questionnaire questions followed the flow-chart process shown in Figure 7.1, with all participants viewing the same information sheet, consent form and end of questionnaire page.

7.3.5 Data Handling

Overall, 256 participants began the questionnaire, but 42 responses were removed for one of the following reasons: no progress past the consent page; being ineligible for participation (such as being an undergraduate student), or due to providing contradictory responses (e.g. replying 'yes' to having used a power analysis, but subsequently declaring that they have never used one in a later free-text box). The final data set consisted of data from 214 participants, details of which are found in section 7.3.8.

7.3.6 Quantitative Analysis

All analyses for this project were exploratory, using the same analysis methods adopted in Chapters 3 and 5. Quantitative analysis took the form of descriptive quantitative analyses and explorations of demographic differences using chi-square tests. Note that the variable Job Role was collapsed into four categories (pre doctoral level participants, doctoral students,

post-doctoral academics and tenure-level staff) for the purpose of these exploratory chi-square analyses, due to small group sizes for expected frequencies. Quantitative analyses were computed using Jamovi (The Jamovi Project, 2021) or Microsoft Excel.

7.3.7 Qualitative Analysis

Qualitative data was analysed using basic content analysis, in the same approach as the earlier free-text data analysed and presented within this thesis. A detailed description of the basic content analysis method can be found in Chapter 3, Section 3.3.6. Explanations for not using, or for not always using an a priori power analysis, and explanations for using post hoc power analyses, were analysed inductively (a bottom-up approach), where codes and categories were identified from the data. In contrast, definitions of *statistical power* were initially analysed deductively (a top-down approach), using the correct definition of statistical power to code each participant definition as either *incorrect* or *shows understanding.*

Incorrect definitions were characterised by describing other concepts, or making clear mistakes, such as "*the size/strength of the effect*", or "*the ability to detect an effect, given the null hypothesis is true*". This deductive coding process was followed by a round of inductive coding for the *incorrect* definitions, to identify the assortment of mistakes made by participants.

A second round of deductive coding was then applied to the *shows understanding* definitions, to code each for the inclusion of the three key elements of power as per the definition provided by Cumming, "*statistical power is the* ***probability*** *of* ***obtaining statistical significance*** *if the alternative hypothesis is true, that is, if* ***there really is a population effect of a stated size***" (2012, p. 322). Each definition received a point for using a term synonymous to probability, another point for mentioning statistical significance or a similar term such as $p < .05$, and a third point for a mention of a *specified* effect or something equivalent such as "*given the alternative hypothesis is true*". For example, this definition would score three points: "*the* ***probability*** *of detecting a* ***true effect of a given magnitude*** *as* ***significant*** *at a given alpha level*". Scoring was deliberately strict with regards to giving points only when a definition mentioned a *specified* effect as opposed to a general effect, as power relates to a specified effect size. For example, "*detect the effect of interest*" or "*an effect of a given size*" would be acceptable, as opposed to the less specific "*the chance of detecting an effect*".

The full analysis of definitions was completed by EC. A random 20% subset was analysed independently by RW to ensure high inter-rater reliability, both for categorising definitions and also for scoring them for mentioning the three key elements mentioned above. Cohen's kappa for categorising definitions was .988, and Cohen's kappa for scoring definitions was .920. Both of these kappa values correspond to 'almost perfect' agreement as suggested by Landis and Koch (1977, p. 165).

7.3.8 Participants

The demographic characteristics of the sample (*n = 214*) are displayed in Tables 7.2 and 7.3. Just under half of participants were PhD students or equivalent doctoral-level trainees, while just over half of participants (54.67%) self-identified as having some kind of involvement with open science or other behaviours connected to improving psychological science. More than two thirds of the sample (68.2%) were from the United Kingdom, with a further 10.3% from the United States of America, and the remaining 21.5% from another 12 countries.

Table 7.2

Demographic Characteristics of the Sample (n = 214)

Demographic Groups	Frequency
Job Role	
Research or Teaching Assistant (no PhD)	**7** (3.3%)
MSc Student	**2** (0.9%)
PhD Student or equivalent trainee	**102** (47.7%)
Postdoctoral Researcher	**23** (10.8%)
Lecturer or Senior Lecturer	**52** (24.3%)
Professor	**15** (7.0%)
Other[a]	**4** (1.9%)
Prefer not to say	**0** -
Missing	***9** (4.2%)*
Location	
Australia	**3** (1.4%)
Belgium	**1** (0.5%)
Canada	**3** (1.4%)
Denmark	**1** (0.5%)
Finland	**1** (0.5%)
Germany	**6** (2.8%)
Ireland	**3** (1.4%)
The Netherlands	**7** (3.3%)
New Zealand	**2** (0.9%)
Saudi Arabia	**1** (0.5%)

South Africa	**2** (0.9%)
Sweden	**4** (1.9%)
United Kingdom*	**146** (68.2%)
England	***81** (37.9%)*
Northern Ireland	***1** (0.5%)*
Scotland	***56** (26.2%)*
Wales	*7 (3.3%)*
"UK" [b]	***1** (0.5%)*
United States of America	**22** (10.3%)
Missing	***12** (5.6%)*
Open Science	
Yes	**117** (54.7%)
No	**84** (39.3%)
Prefer not to say	**3** (1.4%)
Missing	***10*** (4.7%)

[a] Other jobs: assistant psychologist, data scientist, health improvement officer, and trainee clinical psychologist.
[b] One participant wrote 'UK' instead of providing a devolved nation.

Many sub-fields of psychology were represented in this sample. The most common were cognition, health psychology and social psychology, making up 16.82%, 15.89% and 11.68% of the sample respectively. A wide variety of sub-fields such as consumer psychology, environmental psychology, legal psychology and sports psychology were also reported by participants (within the 'other' category in Table 7.3).

Table 7.3
Sub-Fields of Psychology in Sample (n = 214)

Field	Frequency
Behavioural	**2** (0.9%)
Clinical	**16** (7.5%)
Cognition	**36** (16.8%)
Comparative	**6** (2.8%)
Counselling	**3** (1.4%)
Cyberpsychology	**2** (0.9%)
Developmental	**15** (7.0%)
Educational	**4** (1.9%)
Evolutionary	**3** (1.4%)
Experimental	**3** (1.4%)
Forensic	**7** (3.3%)
Health	**34** (15.9%)
Mental Health	**3** (1.4%)

Mathematical	**5** (2.3%)
Metascience	**3** (1.4%)
Neuropsychology	**12** (5.6%)
Occupational	**2** (0.9%)
Personality	**3** (1.4%)
Social	**25** (11.7%)
Other [a]	**13** (6.1%)
Missing	***17 (7.9%)***

[a] 'Other' sub fields, each reported by one participant, were: affective psychology, applied psychology, autism, biopsychology, consumer psychology, cross-cultural psychology, decision science, environmental psychology, legal psychology, moral psychology, music psychology, sexology and sports psychology

7.4 Results

The majority of participants in this sample indicated a belief that power is very, or somewhat important in psychological research (as shown in Table 7.4). Perceptions of the importance of statistical power did not differ significantly by open science category (χ^2 (4, N = 201) = 2.01, p = .0734, V = 0.10), but there was a small statistically significant effect of job role (χ^2 (12, N = 201) = 22.7, p = .03, V = 0.19; breakdown shown in Table 7.4).

Table 7.4

The Importance of Power in Psychological Research (n = 206)

	Perceived Importance of Power				
	Very	**Somewhat**	**Not very**	**Not at all**	**Don't Know**
Full Sample	**127** (61.7%)	**66** (32%)	**5** (2.4%)	**1** (0.5%)	**7** (3.4%)
Job Role					
MSc Student	**2** (1.6%)	**0** -	**0** -	**0** -	**0** -
Research/Teaching Assistant	**4** (3.2%)	**2** (3.0%)	**1** (20.0%)	**0** -	**0** -
PhD Student	**62** (48.8%)	**36** (54.6%)	**0** -	**0** -	**4** (57.2%)
Postdoctoral Researcher	**17** (13.4%)	**5** (7.6%)	**1** (20.0%)	**0** -	**0** -
Lecturer /Senior Lecturer	**28** (22.1%)	**20** (30.3%)	**2** (40.0%)	**0** -	**2** (28.6%)
Professor	**10** (7.9%)	**2** (3.0%)	**1** (20.0%)	**1** (100%)	**1** (14.3%)
Other	**3** (2.4%)	**1** (1.5%)	**0** -	**0** -	**0** -
Missing	***1 (0.8%)***	***0*** -	***0*** -	***0*** -	***0*** -

7.4.1 Part 1: A Priori Power Analysis Use

Self-reported use of a priori power analysis was high in the surveyed sample, as shown in Table 7.5. One hundred and eighty four participants (86%) had experience of using power analysis for sample size planning, although 90 of these 184 participants reported using power analysis alongside other sample size planning methods, such as convenience sampling, or following general rules of thumb for particular research designs.

Table 7.5

Experience using Power Analysis, With Demographic Differences (n = 214)

	Experience Using A Priori Power Analysis	
	Yes	**No**
Full Sample	**184** (86%)	**30** (14%)
Job Role		
MSc Student	**2** (1.1%)	**0** -
Research/Teaching Assistant	**6** (3.3%)	**1** (3.3%)
PhD Student	**79** (42.9%)	**23** (76.7%)
Postdoctoral Researcher	**21** (11.4%)	**2** (6.7%)
Lecturer/Senior Lecturer	**50** (27.2%)	**2** (6.7%)
Professor	**13** (7.1%)	**2** (6.7%)
Other	**4** (2.2%)	**0** -
Missing	***9** (4.9%)*	***0** -*
Open Science		
Yes	**101** (54.9%)	**16** (53.3%)
No	**72** (39.1%)	**12** (40.0%)
Prefer not to say	**1** (0.5%)	**2** (6.7%)
Missing	***10** (5.4%)*	***2** (6.7%)*

The 30 participants with no experience of power analysis ranged from research assistants through to professors, with a small statistically significant difference between job roles (χ^2 (3, N = 201) = 10.2, p = .017, V = 0.22) when jobs were merged into four broader categories. There was no significance difference in power analysis use between those who did or did not report engaging with open science or psychological reform (χ^2 (1, N = 201) = 0.015, p = .902, V = 0.009).

Participants with experience of a priori power analysis (*n = 184*) were asked to estimate the frequency at which they use it, as a proportion of suitable (confirmatory hypothesis testing) studies. Eighty one participants reported using a priori power analysis 100% of the time,

compared to 103 who do not always use it. The overall mean frequency was 79.1% (SD = 27.8), with a median of 90%, and mode of 100%. Estimated frequencies ranged from 9% to 100% of the time.

Software Preferences

Participants with experience of a priori power analysis indicated widespread use of G*Power (Faul et al., 2007), reported 128 times. The second most popular option was R (R Core Team, 2022) (*n = 55*), with the pwr (Champely, 2017) and simr (Green & MacLeod, 2019) packages mentioned most frequently. Eleven participants reported using 'online calculators' without additional detail, and other software choices, each mentioned fewer than five times, were: BrawStats, Excel, Jamovi, JASP, MATLAB, NQuery, PowerPlus, SAS, SPSS, and STATA.

Effect Size Estimation

Methods of effect size estimation for a priori power analyses were varied, with many participants reportedly using multiple approaches. The list of options which was presented to participants is shown in Table 7.6, alongside the number of times each method was selected by participants. The 'Exclusive Use' column highlights how many participants exclusively used that individual method (i.e. did not report using any other approaches to estimation). The most frequently selected method was using an effect size from the results of other published literature, followed by using Cohen's recommendations or similar guidelines. The least popular listed option was asking for recommendations from other researchers (selected 35 times), and only 10 participants used an 'other' method.

Table 7.6

Frequencies for Each Effect Size Estimation Method

	Method	Frequency	*Exclusive Use*
1	Use an effect size from the results of other published literature	122	*9*
2	Use the same effect size as a previous similar study reported in their methods	83	*2*
3	Use a small or medium effect size e.g. Cohen's recommendations	106	*16*
4	Use recommendations from other researchers	35	*2*
5	Use the smallest effect size of interest for my field or "meaningful" effect size for my field	79	*4*

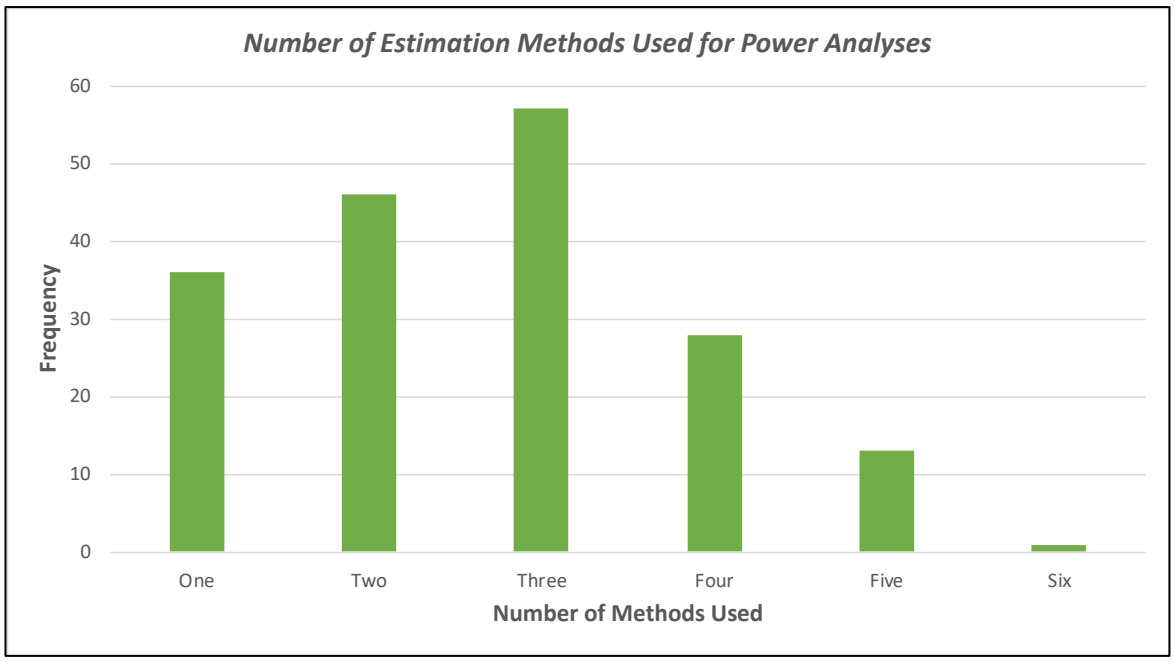
6	Run a pilot study to calculate an effect size first	47	*0*
7	Other	10	*3*
	– *Relying on statisticians to decide*	*2*	
	– *Using scaled-down estimates to account for publication bias*	*2*	
	– *Taking into account sensitivity analyses*	*3*	
	– *Relying on personal unpublished work*	*1*	
	– *Using a personally meaningful effect size*	*1*	
	– *No explanation given*	*1*	

Note. The bottom section of this table lists the 10 'other' methods shared by participants.

As shown in Table 7.6, 36 participants (19.6%) reported only using one method of effect size estimation (as shown in the 'Exclusive Use' column). The majority of participants reported using more than one method for effect size estimation, using a median of three approaches (as shown in Figure 7.2). No participants selected all seven possible options.

Figure 7.2

Number of Effect Size Estimation Methods Used by Participants for Power Analyses



Not Using A Priori Power Analysis

The explanations for not using power analysis shared by those who have never used it (*n* = *30*) and those who do not always use it (*n* = *103*) are presented in Table 7.7. Most frequently, participants chose not to use an a priori power analysis when they knew they would not be able to achieve the large sample sizes it would inevitably suggest. Many other participants

reported difficulties with the calculation or concept itself; either not knowing enough about power in general, or struggling with the specific power analysis calculations. Typical difficulties included a lack of previous literature to use as a guide, or more specifically struggling to determine suitable population effect size estimates.

It should be noted that 12 of these participants clarified that they were taking into account historic behaviour and *do* actually use an a priori power analysis for 100% of suitable recent and future studies, providing explanations such as "*[I] include studies I've done pre-replication crisis*".

Table 7.7

Reasons Why Participants Don't Always Use A Priori Power Analysis

Reason	**Frequency**	
	Power Analysis Use	
	Never	Not Always
Don't know enough about power	6	1
Power analysis is difficult to do	-	4
Unsure about effect size estimation	-	12
Lack of previous literature to inform decisions	-	4
Power analysis is too difficult for complex statistical designs	1	14
Produces unrealistic sample sizes	3	16
Influenced by colleagues	3	6
Influenced by time pressure	-	3
Not needed (no explanation)	3	-
Not needed (not applicable to work)	7	1
Not needed (access to large samples)	1	2
Use other rules and approaches to sample size planning	2	11
Rely on statisticians	1	1
Reflecting on historic behaviour	-	12
Choose not to	1	4
Other	1	6

Both groups of participants had several explanations in common, such as not using an a priori power analysis because it would suggest unrealistic large sample sizes, being negatively influenced by colleagues, and struggling with power analyses for complex study designs such as multi-level models. Several participants commented on using other rules and approaches to sample size planning such as "*[I] knew that as long as I met Tabachnick's and Fidell's rule then I'd be ok*", and the less mathematical "*if the sample size is larger than in comparable*

studies in peer reviewed journals, then I assume that I am safe". Other explanations were tied directly to the calculation itself, most frequently reporting difficulty with effect size estimation, or commenting that power analysis is too difficult (or impossible) for complex statistical designs.

The "other" category in Table 7.7 represents a wide range of responses from participants not using power analysis in 100% of suitable studies, including not using power analyses when working with students, not using a power analysis for direct replications, and preferring to use sensitivity analyses. One participant offered a particularly critical perspective on the use of the power analysis for the sake of journal guidelines, as shown in the quote below:

"The poor understanding of power among co-authors and reviewers is a punishers [sic] *for me to do power analyses well. I've had multiple situations where were people are satisfied with seeing "a" power analysis even though it's wrong. Doing it right can take a lot of effort, and honestly sometimes I wonder why I'm bothering"*.

One other participant, who reported never using power analysis, also criticised the inherent relationship between NHST and power:

"It strikes me that Power analysis is a way of finding what would be a significant value (seeking a p-value)".

7.4.2 Part 2: Post Hoc Power Analysis

All questionnaire participants, regardless of a priori power analysis experience, were asked if they had ever used post hoc power analysis. They were then asked why they had used it if they responded *yes*. Frequencies of post hoc power analysis are presented in Table 7.8, divided into experience of a priori power analysis (*yes* or *no*).

Table 7.8

Experience Using Post Hoc Power Analysis, Divided by Experience of A Priori Power Analysis (Yes or No)

	Used Post Hoc Power Analysis?		
	Yes	**No**	**Missing**
Full Sample	**97** (45.3%)	**110** (51.4%)	*7 (3.3%)*

A Priori - Yes	**92** (94.8%)	**85** (77.3%)	*7 (100%)*
A Priori – No	**5** (5.2%)	**25** (22.7%)	*0 -*

Of the 97 participants who have used post hoc power analyses, 86 provided one or more reasons explaining why they have used the calculation. The most common explanation was simply to check the actual power of a study (e.g. "*to prove that the research was well powered"*), which demonstrates that there are still widely-held misconceptions about post hoc power analysis. A more detailed breakdown of explanations is presented in Table 7.9. The "other" category spans a variety of explanations, including "*when reading the lit other studies that seem rigorous do so*" and "*it is not always clear what the right power analysis is*".

Table 7.9.
Explanations for Using Post Hoc Power Analysis

Reason	**Frequency**
Historic behaviour	11
For educational purposes	3
Personal curiosity	6
Required to do so (publishing or exams)	14
Check actual power	53
- General	*11*
- Due to null results	*6*
- Due to underrecruiting participants	*7*
- Due to secondary data	*3*
- Due to unexpectedly small effect sizes	*2*
- Due to not calculating an a priori power analysis	*5*
- After changing study designs during research	*4*
- In order to demonstrate reliability of findings	*8*
- In order to plan larger future studies	*6*
Calculated for the purpose of a meta-analysis	2
Reproduce calculations when reviewing	3
Other	5

Reassuringly, eleven participants explained their use of post hoc power analysis as being historic behaviour, with several explaining that they had used it before learning about the statistical issues associated with post hoc power. This more knowledgeable perspective was shared by the two participants who mentioned using post hoc power for educational purposes, for instance: *"to demonstrate (using a simulation) to students how crazily it bounces around with replication".* One participant, who explained that they had used post hoc power to

satisfy a reviewer, also commented that doing so went against their personal preferences and that they were aware that it is a nonsensical calculation.

It should also be noted that five participants, not included in Table 7.9, answered 'yes' to using post hoc power analyses, but their explanations indicated that they actually had used a sensitivity power analysis, such as "*tested what the minimum effect I could have detected with my sample size is*".

7.4.3 Part 3: Defining 'Statistical Power'

All participants were asked to define power in their own words, or to write 'I don't know' if preferred. A content analysis of responses identified 57 as *incorrect*, 135 as *shows understanding*, and the remaining 13 were cases of participants stating "I don't know". These results are shown in Table 7.10, subdivided by a priori power analysis experience (*yes* or *no*). Of the 57 incorrect definitions, 40 were provided by participants *with* experience of power analysis.

Table 7.10

Categorisation of Definitions of Power for Full Sample, and Divided by Experience of A Priori Power Analysis (Yes or No)

	Definition Category			
	Shows Understanding	**Incorrect**	**'I Don't Know'**	**Missing**
Full Sample	**135** (63.1%)	**57** (26.6%)	**13** (6.1%)	***9*** *(4.2%)*
A Priori - Yes	**125** (92.6%)	**40** (70.2%)	**10** (76.9%)	**9** *(100%)*
A Priori - No	**10** (7.4%)	**17** (29.8%)	**3** (23.1%)	**0** -

There were no significant demographic differences between job roles and definition categories, χ^2 (6, N = 198) = 8.94, p = .177, V = 0.15. However, there was a significant relationship between open science category and definition category, (χ^2 (2, N = 199) = 12.3, p = .002, V = 0.25), with two thirds of the 'shows understanding' definitions coming from participants who do identify as being involved in open science or psychological reform.

Incorrect Definitions of Power

Several common mistakes emerged in the definitions of power given by participants, full details of which are found in Table 7.11. Some participants clearly defined other statistical concepts, such as incorrectly describing an effect size instead of power; or describing a power

analysis instead of power itself. Three participants also confused power and Type I errors, defining power as "*the likelihood any significant effect is not due to chance*" or similar, while seven participants mistakenly described power as the Type II error rate (as opposed to 1 – the Type II error rate). The most frequent mistake was defining power as a measure of sample size, such as "*the number of participants needed to show an effect*" and "*the minimum sample size needed to be confident that any conclusion drawn is valid*", while a number of other participants provided broader definitions linking power to validity, meaningfulness, and representativeness. The "other" category groups incorrect ideas which only appeared once, such as "*is it about taking the log of a number normalized to a standard value?"* and *"it is like the impact of the finding given the sample and figures*".

Table 7.11

Errors in Definitions of Power, with Frequencies, Divided Into Power Analysis Experience (Yes or No)

Power Incorrectly Defined As…	**Frequency**	
	A Priori – Yes	**A Priori - No**
Effect size	-	2
Power analysis	2	2
Type I error	3	-
Type II error	7	-
Sample size	15	7
- general	*1*	-
- minimum sample size	*6*	*2*
- sample size for meaningful results	*4*	-
- sample size for reliable results	*2*	*1*
- sample size for representative results	-	*1*
- sample size for validity	-	*1*
Measure of meaningful results	5	2
Measure of representative results	3	3
Measure of validity	2	-
Other	4	3

Definitions Rated as "Shows Understanding"

Scores out of three were calculated for all 135 definitions rated as 'shows understanding', as shown in Table 7.12. Most commonly, participants scored two out of three. Scoring was deliberately strict for mentioning a specified effect as opposed to a general effect, as power relates to a specified effect size.

Table 7.12
Scores for Definitions Rated as Shows Understanding

Score out of Three	Zero	One	Two	Three
Frequency	2 (1%)	51 (38%)	65 (48%)	17 (13%)

Seventeen definitions scored three out of three. Examples scored this way include "*the probability of finding a significant effect according to null hypothesis significance testing, given a stated effect size*" and "*the probability that p will be <.05 assuming the alternate is true and a certain effect size, with a given n*".

The two definitions that scored zero but were not classed as incorrect were categorised this way because they indicated some understanding that power relates to the chance of identifying an effect if it exists, but did not quite mention any of the three key elements. For instance, one definition ("*I define power as whether or not my study has the power to detect an effect in the data, if there is an effect to be detected at all*") mentioned *an* effect instead of a specified effect, and uses binary 'whether or not' language instead of referencing probability or an appropriate synonym. This resulted in a zero score but categorisation as 'shows understanding'.

Initial analysis identified 96 out of 135 definitions which directly used the word 'probability' or presented a definition in the format *1 – Type II error* rate, indirectly indicating probability. However, due to the prevalence of other similar terms such as ability and capability, the criteria 'mentions probability' was expanded to include the mention of any synonyms, increasing the frequency to 126 out of 135. All definitions mentioning a similar term were scored as mentioning probability.

Sixty-six out of 135 definitions mentioned statistical significance, or provided similar descriptions such as correctly rejecting a false null hypothesis, or power being associated with a set alpha. Example definitions which were scored as mentioning statistical significance (or describing the same concept) include "*the chance of an effect to be detected (according to a set alpha) given the effect is true*" and "*the ability to detect a (statistically significant) effect*".

Only 37 out of 135 definitions mentioned a particular effect (as opposed to using general language about effects existing) and therefore were scored as correctly describing this third element of power. For instance, a definition such as "*the probability of detecting the effect you have predicted, assuming that is the true effect*" correctly refers to a specific effect size and was scored as mentioning this element; compared to "*the ability to detect a (statistically significant) effect*", which only refers to an unspecified effect. However, taking into account all mentions of 'an effect' and similarly vague terms, 101 participants made some reference to effects, and two more mentioned finding a '*significant difference*'. Further analysis looking for mentions of an effect 'truly existing', or other similar language, found that 75 definitions clearly stated that an effect needed to exist or was real, such as "*the ability to identify an effect if an effect truly exists in the population*". An additional 15 participants referenced a 'true alternative hypothesis' or 'false null hypothesis'.

7.5 Discussion

Despite widespread criticism, NHST remains the dominant approach to data analysis in psychology. In order to improve its use, more attention is being paid to statistical power, which aims to reduce the uncertainty associated with NHST by reducing the chance of a Type II error. The evaluation of power is advocated for by large organisations such as the APA (e.g. Appelbaum et al., 2008), and as seen in Chapter 2, many journals now encourage the use of power analyses for sample size planning. However, recent research demonstrates that researchers lack intuitions about power (Bakker et al., 2016) and that encouraging authors to use power analyses does not necessarily result in them being computed correctly (Bakker et al., 2020). This chapter has made a novel contribution to the literature by presenting new data related to the use and knowledge of statistical power in the psychology researcher population, including qualitative insights into individual experiences.

7.5.1 Experiences Using Power Analyses

Within the current study, self-reported use of a priori power analyses was notably high, with just 30 of 214 participants having never used it. However, similarly to Bakker et al. (2016), a high proportion of participants acknowledged that they do not use power analyses for all suitable studies, for a variety of reasons. The most important barrier identified in this study is difficulty with the actual process of calculating power, including difficulties with software,

struggling to determine suitable effect sizes, or struggling with power for complex research designs. Given that the majority of participants report using G*power, which uses lots of mathematical language throughout its manual, and does not support multi-level regression or similarly complex models, perhaps it is unsurprising that so many report these problems. As alternatives to G*power grow in popularity (such as apps built in R), it is likely that more user-friendly interfaces, and more context-specific calculators will become available, which will help researchers overcome barriers related to software and calculations. However, it will take more work than this to overcome issues related to effect size estimation, as discussed below.

7.5.2 Understanding Power and Effect Sizes

Statistical knowledge is important from two different perspectives within this topic: direct knowledge of statistical power itself, and also knowledge of effect sizes, given that effect size estimation is crucial to the power analysis process. While a lack of knowledge about power was only infrequently mentioned by participants as an issue, other data reported here indicates that many researchers lack knowledge related to power. For example, despite the growing criticisms of post hoc power discussed in Section 7.2, many participants still believe that post hoc power is an appropriate way to calculate 'actual' study power. Concerningly, several also commented that they were asked to compute post hoc power by editors or reviewers, indicating a lack of knowledge within the community that is responsible for the quality of published literature. Beyond this, many of the incomplete or confused definitions of power suggest that few researchers grasp that power is both tied to finding a statistically significant *p*-value, and tied to a population effect size.

If researchers only have an insecure understanding of what power actually is, they are unlikely to input sensible values into an a priori power analysis calculation (where a 'sensible' value would be an effect size estimate which does not overestimate the population effect size). Overestimating the population effect size corresponds to underestimating suitable sample sizes; and so does not actually decrease the risk of Type II errors. Similarly, without understanding what power is, researchers are unlikely to adequately discuss or critically evaluate the power of their study, as requested by journals who adhere to the JARS guidelines (Appelbaum et al., 2018). The lack of knowledge of others also risks demotivating those who do try and improve their own methods, as expressed by one participant who commented on being discouraged by reviewers who are often satisfied with seeing any power analysis, even

if it's wrong. This broader lack of knowledge turns power analyses into a new tick-box exercise, instead of an educational tool to use within the research cycle.

Knowledge of effect sizes is also crucial within the domain of statistical power. Chapter 3 demonstrated that many researchers have a limited knowledge of effect sizes, and the data presented in this chapter suggests that researchers are equally uncertain when estimating effect sizes for power analyses. Many researchers in this reported still using Cohen's guidelines, which have been criticised for lacking specificity and relevance to each field (e.g. Correll et al., 2020). Of course, while some researchers may be entirely aware of their limitations yet use these benchmarks anyway, it could also be that knowledge of their issues is not yet widespread. The only estimation approach used more frequently was to take effect sizes from the results of previous literature, which is also problematic as historic literature is likely to overestimate effect sizes due to a combination of small sample sizes and publication bias (see Smaldino & McElreath, 2016; Simmons et al., 2011). As discussed above, if inflated effect sizes are used in future power analyses, suggested sample sizes will remain smaller than necessary, and Type II error rates are unlikely to decrease. The self-reported behaviours found in this study align with reviews of the literature, such as recent research by Bakker et al. (2020), who found that when reported power analyses did explain their effect size choices, they often referred to Cohen's guidelines or past literature to justify their choices. It is apparent that researchers need to be better-equipped with tutorials regarding effect size estimation, to ensure they are using appropriate and consistent approaches in their calculations.

7.5.3 Study Limitations

The primary limitation of this study is an overall lack of generalisability to the wider psychology researcher population, similar to the previous questionnaire studies reported in this thesis. For example, self-reported power analysis use is very high; which may reflect the growing requirements for power analysis use within journals, but could equally be an overestimate of current behaviour due to self-report biases. Given that the review shared by Tressoldi & Giofrè (2015) found power analysis reporting to be as low as 3%, it is unlikely that, just a few years later, true rates of power analysis use in the wider psychology community are as high as the 71% measured in the present study. However, it could also be attributed, in part, to the high proportion of participants in this study who engage with some aspect of psychological reform or open science, who may be more likely to think critically

about power and sample size, and adopt behaviours such as power analyses. The sample also heavily features PhD students and early career researchers, who are more likely to have only been involved in psychology since the replication crisis and subsequent statistical reform period, and hence could have been exposed to discussions of power and sample size throughout the majority of their careers.

7.5.4 Future Directions

While broad knowledge of power is useful to ensure that participants are capable of critically evaluating the power of their work, effect size estimation is particularly important as overestimates of a population effect size will result in underestimates of suitable sample sizes. Future research could expand on this particular topic by asking participants to describe their estimation decision process (instead of a multiple-choice list) to more accurately capture their behaviour, or could ask participants to choose their most common method if using more than one. From a wider perspective, future research could also take the form of reviews, such as expanding the work of Bakker et al. (2020) to examine the contents of reported power analyses, or to explore whether encouraging or requiring power analyses or power discussions has a positive influence on reporting behaviour.

The strict approach to banning post hoc power employed by two journals in Chapter 2 is something that could be valuable to adopt across all journals, as it is apparent that many researchers in psychology have not yet discovered that post hoc power analyses are generally uninformative. Journals may also be an accessible location for educational resources and power analysis apps (although in Chapter 2, very few of the reviewed journals made resources available to authors). Similarly, given their accessibility and influence, organisations such as the APA could expand their reporting standards to include educational materials and clearer guidance for power analyses. Indeed, funding bodies could be similarly proactive in providing educational support to researchers, particularly as many funding applications now require power analyses to be included.

7.6 Conclusion: The Messy Reality of Power

Many would argue that the ideal psychology researcher would in fact not require any knowledge of statistical power, as the ideal researcher would not use the NHST framework at

all. This is a perfectly reasonable argument, given the criticisms of NHST that have been acknowledged earlier in this thesis. However, NHST remains prevalent in psychology, and is unlikely to simply disappear. And so, the ideal researcher should have a firm grasp of what statistical power is, be able to calculate power analyses using appropriate estimates of effect size, and dismiss post hoc power due to its overlap with *p*-values. In addition, they would be able to critically evaluate statistical power within a wider context of sampling error and other limitations, rather than perceiving high power as a guarantee of accurate results.

The data presented in this chapter highlights the messy reality of psychology researchers who are increasingly being expected to incorporate statistical power analyses into their work. Knowledge of statistical power, and in particular post hoc power, is broadly poor, which will likely correspond to the lazy justification of 'low power' to explain away null findings, or 'high power' to give excessive confidence in significant findings. Only those equipped with a firm grasp of statistical power will understand that the Type II error rate is still unknown, even if a power analysis calculation uses a particular power level. More broadly, focusing too heavily on power also fails to acknowledge Type I errors, which will always exist regardless of sample size. The data shared here highlights that power analyses risk becoming a new tick box exercise, rather than a way to exercise any critical thinking before, during, or after a study has been carried out – particularly given that this messy reality appears to be true of editors and reviewers too.

Chapter 8: General Discussion

Effect sizes, confidence intervals and statistical power are three concepts that are often promoted within discussions of statistical reform. While none of these statistical concepts are technically new, focus on their use has increased in recent decades, in part as a response to the issues highlighted by the replication crisis. However, to be effective in improving the quality of psychological science, each must be used appropriately and purposefully, and understood well, to avoid repeating past mistakes made with NHST. This chapter will review the objectives and findings of this thesis, highlighting the novel contributions it has made to the literature. This will be followed by a discussion of the thesis limitations, and will conclude by reflecting critically on the use of effect sizes, confidence intervals, and power within statistical reform.

8.1 Thesis Objectives & Findings

1. To review the current contents of psychology journal author guidelines to identify the presence of any statistics guidelines, particularly looking for comments on NHST, or the inclusion of (1) effect sizes, (2) confidence intervals or (3) statistical power.

2. To examine how frequently psychology researchers report using each of the three aforementioned statistical concepts, along with their explanations for not using them.

3. To examine knowledge, understanding and interpretations of each statistical concept, identifying the prevalence of any misconceptions.

8.1.1. Objective 1: Journals and Statistical Guidelines

Objective 1 corresponds to the review presented in Chapter 2 of this thesis. This review makes a novel contribution to the literature as one of the first to examine and report on the statistical guidelines of a large sample of psychology journals. Despite the widespread criticism of NHST discussed in Chapter 1, few journals had any noticeable position on the use of NHST. Typically, advice related to NHST was limited to asking authors to report *p*-values using exact numbers, or consisted of vague comments such "*avoid relying on p-values*", which is found within the ICMJE's guidelines (ICMJE, 2021, p. 17). Just three

journals provided more explicit advice related to *p*-values: one asking authors to correct for multiple testing, and two Springer journals with in-depth guidance on best practices for working with NHST.

In contrast, the findings reported in Chapter 2 highlight that effect sizes and confidence intervals are now widely encouraged across psychology. Seventy-four of the top 100 journals included instructions related to effect sizes, compared to 68/100 who included instructions related to confidence intervals. These findings align with the published literature, where effect sizes appear more frequently than confidence intervals (e.g. Fritz et al., 2013). However, efforts to encourage authors to make use of these statistics to evaluate their data were similarly poor for both, with just three journals asking authors to discuss effect sizes, and zero asking authors to discuss their confidence intervals. Given that a lack of interpretation is widely noted across the literature (e.g. Fidler et al., 2004; Peng et al., 2013), future updates to author guidelines could incorporate more guidance on making use of reported statistics, rather than handling statistics primarily as a tick-box exercise. However, as confidence interval interpretation is mired in statistical controversy, with some advising that a single confidence interval is perhaps best not interpreted at all ("*how does one then interpret the interval? The answer is quite straightforward: one does not*" (Morey et al., 2016a, p. 118)), it is hard for journals to advise authors on how to interpret confidence intervals. Nonetheless, broader requests for authors to reflect on the uncertainty of their findings would circumvent this conflict, while still adopting the recommendations of the estimation approach.

More than half of the top 100 journals asked for sample sizes to be justified, although not all specifically referenced statistical power. Typically, power analysis was encouraged rather than required. This is sensible given that statistical power is tied to the NHST framework, which some researchers prefer to avoid entirely (e.g. using Bayesian options instead). However, three journals required a priori power analyses or sensitivity power analyses, which appears to force all submissions to remain tied to NHST to some extent. Of course, using a power analysis does not force a researcher to also use NHST for analysis, but requirements associated with statistical power fuel the narrative that significant *p*-values are still the primary desirable outcome for research. Reflecting on the problems related to post hoc power discussed in both Chapter 1 and Chapter 7 of this thesis, it was also noted that just two journals have explicitly banned post hoc power calculations from submitted work.

8.1.2 Objective 2: Using Statistics in Psychology

The three questionnaire studies reported on in Chapter 3, Chapter 5 and Chapter 7 of this thesis collectively fulfil Objective 2, providing insight into the use of effect sizes, confidence intervals and power analyses within psychology. Minimal research exists into the use of these three statistics from individual perspectives, with the literature typically focusing on reviewing reporting practices across published articles. As such, the studies in this thesis offer novel insights into the individual experiences of psychology researchers. Use of all three statistics was high across the respondents studied here, with just 10% of participants never using effect sizes, 10% of participants never using confidence intervals, and 14% of participants having no experience of using any power analyses.

With regards to effect sizes, a further 20% acknowledged 'not always' using effect sizes when reporting quantitative data, while the remaining 70% of participants claimed to always use them. However, a particular issue when considering the use of effect sizes data is that further data collected in Chapter 3 indicates that many researchers may only be aware of a very limited number of effect size measures, or only know of a few situations where effect sizes can be used (e.g. many referred to an effect size always measuring the difference between two groups). Taking this into account, it is likely that use of effect sizes has been overstated by participants, if they are imagining 'use' only within the limited contexts that they know of.

Confidence intervals had a more balanced set of responses, with 41% of participants not always reporting them with quantitative research, compared to 49% of participants who reported always using them. Similarly, while 86% of participants had experience using power analysis, they were not used for all research. On average, participants reported using them for 80% of their quantitative work (with a range from 9% to 100%). Power analyses were also typically not used exclusively for sample size planning, which is similar to other findings in the literature (e.g. Bakker et al., 2016).

Several explanations for not (or not always) using these statistics recurred across all three concepts. For instance, lack of knowledge of each particular concept was widely stated as a barrier to use, which further justifies the subsequent data collected as part of this PhD to identify knowledge gaps and misconceptions. In the case of both effect sizes and power,

several participants also commented on a collective lack of knowledge as a further barrier, such as a lack of consensus on which effect size indices are suitable for complex models, or a lack of tutorials explaining how to do a priori power analyses for similarly complex studies. The role of requirements was also notable across all three questionnaires, with many participants acknowledging that their behaviour is related to the requirements (or lack of) in any given situation. Many of these participants explicitly connected their statistical reporting to journal requirements, while several others admitted that their own bad habits influenced their decisions – implying that they could also be extrinsically motivated by measures such as journal guidelines. Another source of extrinsic influence which appeared frequently was 'people', mentioned within several different contexts. Many participants admitted that their colleagues or supervisors negatively influenced their statistical habits. Others spoke of 'people' on a broader scale, pointing out that these statistics are not always used in the literature that they consume, and so they do not use them in their own work.

8.1.3 Objective 3: Knowledge, Understanding and Interpretations

The three questionnaire studies reported on in Chapters 3, 5 and 7 also contribute to Objective 3, by demonstrating how well (or poorly) psychology researchers understand effect sizes, confidence intervals and statistical power. This objective was further investigated through the studies in Chapter 4 (effect size visualisation) and Chapter 6 (confidence interval interpretation).

8.1.3.1 Effect Sizes

As there is very little literature on effect size knowledge, Chapters 3 and 4 make a novel contribution to the field by offering an assortment of quantitative and qualitative insights into effect size knowledge and perception. In Chapter 3, performance on a novel true-false knowledge test was strong, with more than three quarters of responses to each statement being correct. If the ideal researcher should have a firm basic knowledge of effect sizes, then this sample appears to come close to meeting this criteria. For instance, very few participants appear to hold the *magnitude fallacy* (the misconception that significance is associated with larger effect sizes), as tested using statements including "*A small effect size indicates that the null hypothesis should fail to be rejected*". However, in contrast, the definitions of effect size shared by participants indicate a messier reality of effect size knowledge, which appears to be limited to very specific contexts (with overly-specific definitions including "*effect sizes*

reflect the magnitude of any given difference between two groups" and "*the size of the change in the outcome following an intervention").* The novel graph study in Chapter 4 further highlights this messy reality of effect size judgement, as there appears to be a disconnect between actual data and written effect size values. The data reported in this thesis indicates that researchers consistently underestimate effect sizes when inspecting raw data, suggesting that researchers have higher expectations of how visible an effect should be than how it appears in reality.

8.1.3.2 Confidence Intervals

While several studies now exist looking at confidence interval knowledge and understanding, the research presented in Chapter 5 makes a novel contribution to the literature on knowledge by using a new true-false scale. This new scale was designed to identify misconceptions instead of focusing on probability (e.g. Hoekstra et al., 2014; Lyu et al., 2018). Responses to the true-false scale indicate that researchers may have a reasonable knowledge of confidence interval widths, but lack knowledge related to samples versus populations, and the association between replications and confidence intervals. In addition, Chapter 5 also presents additional novel data in the form of qualitative definitions of the term confidence interval, with participants offering a wide variety of suggestions ranging from accurate ("*if we were to repeat this same procedure infinite times, 95% of the time the confidence interval constructed would contain the true population parameter the procedure is estimating"*) to clearly incorrect (*"we are 95% confident that we would obtain the results observed here if the null hypothesis is false"*). Overall, the data within this chapter reinforces the findings in the wider literature that confidence interval knowledge is a particularly messy reality, given that no true-false item had more than 64% correct answers, and more than half of definitions were scored as incorrect even when evaluated using the flexible '95% confident' approach to confidence intervals.

This is subsequently further supported by Chapter 6, which makes an additional contribution to the confidence interval literature by providing a contemporary replication and extension of older research using a new psychology researcher sample. If the ideal researcher is one who can interpret confidence intervals to evaluate data with an 'uncertainty' or 'estimation' mindset, this certainly is not yet true in reality, particularly given that many participants still

demonstrated a fixed mindset (‘*this effect exists/does not exist’*) when interpreting a confidence interval.

8.1.3.3 Statistical Power

The data in Chapter 7 makes a novel contribution to the power literature by testing knowledge of statistical power with qualitative data, in contrast to previous research where participants were asked to identify the correct definition of power from a list (Bakker et al., 2016). The definitions of statistical power shared by participants in this thesis reinforce Bakker et al. (2016)’s conclusions that statistical power does not appear to be well-understood. While an ideal researcher should understand that statistical power is associated with statistical significance and an accurate assumption of a population effect size, the messy reality is that many definitions did not refer to NHST (e.g. statistical significance, rejecting the null hypothesis, or a set alpha) at all. Many others either made no mention of an effect size, or vaguely referred to ‘an effect’ instead of specifying that a population effect size is an important component of power. Many participants appear to view power through the lens of sample size planning, confusing the concept of statistical power with power analysis calculations. In addition, participants mistakenly described power as a measure of meaningfulness, representativeness or validity, offering new insights into misconceptions that exist in the psychology researcher population. The data presented in Chapter 7 also reinforces the messy reality of effect size knowledge which has been identified in Chapters 3 and 4, as difficulties with effect size estimation for power analyses were frequently mentioned by participants.

8.2 A Broader Messy Reality: Statistical Reform in Psychology

Thus far this thesis has explored effect sizes, confidence intervals, and statistical power as three key concepts within statistical reform. The data shared by participants throughout this thesis has indicated a messy reality across individuals, with highly varied levels of knowledge, and an assortment of explanations given for not adopting particular statistics at the individual level. However, this messy reality can additionally be seen through a much wider lens, when critically reflecting on each of these three statistical concepts and their value to psychological research.

8.2.1 Effect Sizes

It is difficult to conceptually criticise effect sizes. Given that the term 'effect size' covers all kinds of standardised and unstandardised values, they can be seen as useful in most research contexts. However, it is particularly important to acknowledge that the ideal use of effect sizes is by *interpreting* them to critically evaluate findings. This is what encourages researchers to move away from dichotomous thinking and adopt a more informed critical mindset. In spite of this, in Chapter 2, just three journals asked authors to discuss their effect sizes; and the wider literature demonstrates that effect sizes are often unexamined in any kind of detail (e.g. Peng et al., 2013). Similarly, the APA includes effect size *reporting* within its JARS guidelines, but also fails to include *interpretation* or discussion, despite their Task Force report suggesting that "*it helps to add brief comments that place these effect sizes in a practical and theoretical context*" (Wilkinson, 1999, p. 599).

8.2.2 Confidence Intervals

The obvious criticism of confidence intervals is how difficult they are to interpret, considering that it is mathematically incorrect to explain that one interval has a '*95% chance of containing the true population value*'. These interpretative difficulties are clear to see in the data presented within this thesis, as well as across the wider literature. Indeed, as it is difficult to interpret a single interval and use it to make inferences about the population, its value can be hard to define. In addition, as they are often promoted as an alternative to NHST, or at least as an accompaniment, they should be treated as an independent source of inferential information. However, given that they are often interpreted using NHST logic (e.g. using the presence of zero in an interval as justification for not rejecting a null hypothesis), they do not appear to be viewed as independent by researchers.

However, if confidence intervals are used more generally as an illustration of uncertainty, then their value is more obvious. Within the estimation approach, they are promoted as an answer to the question 'how certain?', which is a question often neglected when evaluating *p*-values. Acknowledging uncertainty is a nudge for researchers to think more critically about *minimising* uncertainty, which may have a knock-on influence on thinking more carefully about research design and variable measurement. Consequently, confidence intervals may offer value even without trying to read them as a specific range of population values.

There is an alternative to the frequentist confidence interval, which *does* allow claims of probability and inference to be made: the Bayesian credible interval. A credible interval can be interpreted as having a particular chance of containing the true population value, and therefore appears to be more useful than a confidence interval (e.g. Morey et al., 2016a). Indeed, credible intervals are also now acknowledged by proponents of the New Statistics as another way to adopt the estimation approach (Calin-Jageman & Cumming, 2019). However, as briefly discussed in Chapter 1, moving to Bayesian statistics requires a new theoretical framework, which is neither used or taught as frequently as the more traditional frequentist approach. While credible intervals may be easier to interpret, they also require even more training and support than adopting more familiar statistics such as confidence intervals. Their similarity may present a useful opportunity to teach students about the frequentist and Bayesian approaches to statistics, which may be a fruitful long-term improvement for psychological research (a suggestion made by Hoekstra et al., 2018). However, given that employed researchers are situated within a publish or perish culture, with time as a common barrier to personal development, recommending that researchers to transition to a Bayesian focus is realistically too great a demand.

8.2.3 Statistical Power

When considering that much of the replication crisis and reform literature focuses on false positive (Type I errors) (e.g. Ioannidis, 2005), while statistical power corresponds to the false negative (Type II error) rate, the relevance of power within statistical reform is easy to question. This is further compounded by arguments to replace the NHST framework, or at least drop the concept of statistical significance, which render power even less useful given that statistical power is tied to the probability of a statistically significant outcome. Indeed, focusing on statistical power arguably continues to over-value significant *p*-values, which detracts attention from wider issues of measurement and design. Power also cannot minimise or detect Type I errors, but increases the potential for a researcher to claim that their study is well-powered and therefore misguidedly claim that their statistically significant finding is 'true'.

However, as NHST remains prevalent in psychology, it is unsurprising that aiming for higher power has become a particular strategy to increase the reliability of science. As the Type I error rate is fixed by a chosen alpha (conventionally 5%), trying to reduce the type II error risk is a sensible precaution for researchers to take, particularly as it is likely to encourage

researchers to collect data from larger (and therefore more representative) samples. Not only does this increase the broader reliability of a piece of work by reducing sampling error, but this also represents a more ethical use of participant time. Perhaps the biggest concern is that focusing on statistical power itself is likely to become a tick-box exercise, given that an a priori power analysis requires assumptions to be made about effect sizes. As shown throughout this thesis, researchers struggle both with effect size estimation for power, and indeed the concept of effect sizes in general.

Some researchers, including those who advocate for the estimation approach discussed within this thesis, propose that planning should revolve around *precision* instead of power (e.g. Cumming, 2012). Typically, the precision approach involves planning a sample size based on the desired width of the resulting confidence intervals (although other less-common strategies exist). One strength of precision is its theoretical independence from NHST, as its goal is to reduce the uncertainty of a set of results, instead of obtaining statistical significance. Advocates also suggest that precision requires less knowledge than power (when used for sample size planning), as it does not require the estimation of a population effect size (e.g. Kelley et al., 2003). However, the confusion over confidence intervals reported both in thesis and elsewhere suggests that moving further towards a confidence interval approach is something that should be handled with caution.

8.3 Key Future Directions and Difficulties

This thesis inspires a number of potential future research projects that could continue to explore statistical reform, journal standards and effect sizes, confidence intervals and power both within and beyond the psychology research population. However, future directions are not limited to just research: there are many ways that change could be implemented across the discipline to enable successful statistical reform.

8.3.1 Future Research Suggestions

With regards to research, the review reported in Chapter 2 could be expanded to review the actual reporting practices found in journals that have each type of guidelines (requirements, recommendations or mixed). This would offer some insight into how much extrinsic motivation researchers require in order to expand their statistical behaviour, particularly when

looking at whether recommendations (without requirements/penalties) still correspond to increased use of statistics. This data could be complemented with interviews of editors and peer reviewers, both to find out more about their attitude towards statistical guidelines, and to identify the extent to which they monitor or enforce statistical reporting behaviours. Given that one participant in Chapter 7 commented that *any* power analysis would satisfy reviewers, even if it is incorrect, it would be interesting to examine this alongside the viewpoints of actual reviewers and editors.

Collectively, Chapter 3, Chapter 5 and Chapter 7 provide important individual perspectives on statistical use, including barriers to use. Future work could offer updated reviews of the literature to examine more recent use of each statistic, given that Fritz et al., (2013) provides one of the most recent large-scale reviews. In addition, contemporary reviews should go beyond basic reporting patterns and establish the extent to which alternative statistics (i.e. non-NHST) are used for *interpreting* data and drawing conclusions; as the value in these statistics is using them for improved evaluation, not just reporting them.

Finally, the assortment of studies from Chapter 3 through to Chapter 7 emphasise the varying levels of knowledge and understanding of statistics within the psychology researcher population. While future research could explore basic and conceptual knowledge further, the most valuable research should focus on interpretation and deeper understanding, similar to the review-style research proposed above. Studying effect size interpretation and deeper understanding could take many forms, including asking participants to compare the results from published studies to establish their approaches and judgements, or investigating the use of standard benchmarks such as Cohen's. The study shared in Chapter 4, for example, could be inverted by asking participants to match a written effect size to one from a small selection of graphs. In contrast, the interpretation study of Chapter 6 has already used perhaps the most obvious way to test interpretation, using written scenarios and qualitative responses. Future work in this specific area could offer more contextual scenarios to participants, which correspond to actual research that they are likely to encounter. Lastly, while basic knowledge of power could be more widely explored, perhaps the most valuable focus for future research is to explore effect size estimation in more detail, to establish how best to support researchers who struggle with a priori power analyses.

8.3.2 Further Actions and Difficulties

Beyond research, various future steps could encourage successful statistical reform across the discipline. The most important is arguably wider education and statistical support, given that both self-reported barriers and the wider data collected in this thesis all highlight individual knowledge as a reason to not adopt various statistics. However, as Chapter 3 and Chapter 5 reported that workload and time are two immediate barriers to training, there is little value in designing materials that will never be used. The most valuable method to support learning would consist of high-level departmental or organisational changes which ensure that researchers have protected time for personal development and training. As this is a particularly ambitious goal, the most practical future direction would simply be to offer materials or training which are short, easily accessible (e.g. online and asynchronous), and context-specific to allow researchers to quickly translate them into their own research.

Similarly high-level changes could also take the form of increased journal recommendations or requirements related to statistics, given that participants across all three questionnaires in this thesis noted that requirements (or a lack of requirements) influence their behaviour. Indeed, journals could also be a useful gateway to education, given that their author guidelines will be accessed by anyone attempting to publish in a particular outlet. However, not only does this require top-level editorial changes, it is also not yet known whether researchers would make use of journal-hosted support. To investigate this further and make evidence-based suggestions, it would be valuable to find out how many researchers read and make use of the information about NHST shared by the Springer Psychonomic journals, or use the New Statistics tutorials shared by Psychological Science (as discussed in Chapter 2).

While both of these suggested higher-level changes have the potential to positively impact how researchers make use of statistics in their work, the reality is much more complex. Firstly, there are basic practical questions that must be considered, including: who is capable of creating accurate yet accessible materials, how will these individuals be identified, and what makes it worth their time and effort to contribute? These practical barriers also apply to journal changes: for example, who is supporting editors if they make high-level changes? Which changes should they make? And what incentives do they have to use further time and effort to enforce them across submissions?

At the more conceptual level, how can educational materials even be produced at all, given that concepts such as effect sizes and confidence intervals lack the more ‘concrete’ rules that

are associated with null hypothesis testing? Null hypothesis testing is, at face level, quite simple to understand: $p < 0.05$ means statistically significant, which offers a simple rule for researchers and learners to grasp (note that this is not necessarily a good thing). In contrast, while an effect size is simple by definition, the reality is far more complex as there are almost infinite possibilities when considering all of the potential standardised and unstandardised ways to measure magnitude. Educating researchers on using effect sizes, where 'any measure of magnitude' counts, is much harder than asking them to make use of familiar and simple *p*-values. Similarly, while effect sizes produce measures of 'size', there are no straightforward sets of guidelines to provide clear and usable interpretations of single values. As such, there is no easy way to answer the question of 'how *big* is $d = 0.3$?', to explain to researchers how to make use of any reported effect sizes to explain the phenomena that they are studying. Confidence intervals are no more concrete, given the disagreement that exists regarding confidence interval interpretation – if we ask researchers to use confidence intervals, but can't provide any concrete explanation of what their single intervals mean, it is hard to demonstrate their value. These complexities mean that not only are these statistics difficult to use, they are also difficult to suitably teach as they do not offer simple rules that can be followed to make judgements about data – particularly within the constructs of time and accessibility that have previously been highlighted by participants in this thesis.

The role of journals is equally complex. For instance, what benefit is there to often-unpaid editors and reviewers who have to take the time to check whether authors have followed statistical guidelines? And who checks if they have sufficient knowledge to do so correctly? Participants in this thesis explicitly reported that 'any power analysis' seems to be enough for reviewers, which de-incentivises researchers to make any effort to use statistical tools correctly. It should also be noted that the broadness of the range of possible effect size indices is often neglected, which was clear in the narrow ways that participants often defined 'effect size' in Chapter 3. If editors or reviewers hold these same narrow views, there is a risk that researchers who make use of more unique or complex and context-specific measures of magnitude may find their work rejected for not meeting particular guidelines. Equally, this could happen in reverse, where researchers begin to feel obliged to provide more common standardised measures of effect size just to make their work 'acceptable' for publication, despite these measures not suiting their data.

Furthermore, despite many participants in this thesis arguing that they are motivated by external sources such as journal requirements, there is certainly another cohort of researchers who are not, given that adherence to statistical requirements is not particularly high in reviews of the published literature (e.g. Giofrè et al., 2017). This could happen for a number of reasons, including a lack of reviewer or editor oversight (perhaps due to time constraints or individual lack of knowledge), lack of knowledge on the part of the researcher, or it could also be attributed to making educated decisions about using other methods of statistical analysis which are less common but no less rigorous. This particular argument raises a further important question to consider: to what extent *should* academics be pushed into changing their statistical practices? And who has the right to pick and choose which statistics are 'best', or indeed try and influence the behaviour of others in the first place? There are evidence-based arguments for the adoption of the estimation approach, for instance, given that it tackles some of the limitations of relying on *p*-values alone, but as discussed throughout this thesis, there are many difficulties associated with the use of effect sizes and confidence intervals. Similarly, there are also perfectly plausible alternatives, such as Bayesian statistics, which have not typically been adopted at the organisational level. Is it fair that researchers are pushed into satisfying particular statistical requirements set by higher-level 'others', even if they are perfectly well-informed on their own statistical practices and are capable of making other choices? Statistical reform efforts must find a way to strike a delicate balance between encouraging evidence-based ways to improve the discipline, without over-policing individuals who are well-equipped to make their own decisions.

8.4 This Thesis as a Case Study: Reflecting on Statistical Practices

This thesis has focused on the use and understanding of effect sizes, confidence intervals, and statistical power within psychological science. However, of these three statistical concepts, only one appears in this thesis: effect sizes. As justified in earlier chapters, power analyses were not deemed appropriate for sample size planning, given that the nature of all work presented here is wholly exploratory. Similarly, confidence intervals do not feature in this thesis. This is partly due to the controversy around their interpretation which has been discussed in this chapter, and is also because the objectives of this thesis do not relate to trying to estimate the uncertainty of any particular finding. These choices could be seen as contradicting some of the ideas discussed in this thesis, such as suggestions that statistical

behaviour change should perhaps be enforced as mandatory by organisations such as journals. However, the decisions here perhaps reflect the optimum approach: to make educated choices about which statistics best suit the work being produced.

While there is not a large number of effect sizes reported through this body of work, Chapters 3 – 9 all report at least one effect size, all used to provide brief comparisons between demographic groups for particular variables. However, despite the emphasis in this thesis (and in the wider literature) on the value of *interpreting* effect sizes, the majority of the effect sizes presented here have not been discussed. This is something that provides a clear illustration of the difficulties faced by the statistical reform movement: is it really sensible to *enforce* the use or interpretation of any given statistic, and force researchers to conform to a single set of rules, regardless of the intentions of a piece of research? The effect sizes reported in this work do not relate to the actual goals of this thesis: they are merely small explorations of the data, included to offer as much transparency as possible to an interested reader. Their inclusion was an informed choice, presented to offer a reader a more well-rounded set of findings than just reporting *p*-values. Similarly, the lack of interpretation of the majority of these effect sizes was also an informed choice, also based on their lack of relevance to the overall goals of the work.

This decision was further reinforced by some of the limitations of the data itself: the demographic groups themselves were consistently very unbalanced in size, and the choice of wording for these questions (which is discussed further in Section 8.5) was less informative than it could have been. Subsequently, interrogating the meaning of any particular effect size related to these variables does not necessarily offer a way to draw valid conclusions.

8.5 Thesis Limitations

The most important limitation of the work presented in this thesis is the generalisability of the findings. Across all studies, there were high proportions of early career researchers, and also researchers who self-identified as engaging with some form of open science behaviour (ranging from 47% of participants in Chapter 3, to 73.2% of participants in Chapter 4). Overall, this thesis lacks sufficient representation of experienced researchers, such as professors or equivalent tenure-level academics; although it is unknown how this might

influence the results. For instance, more experienced researchers could be better equipped with statistical knowledge through longer experience and exposure to research; or conversely, could be more firmly rooted in traditional NHST practices based on what has been familiar throughout their careers. In a similar manner, the high proportion of 'open science' researchers (in all their possible forms) may not be problematic: the key issue here is that it is impossible to estimate how many researchers now engage in at least one type of open science behaviour in order how to determine how representative this data is.

Sample sizes were reasonable, with each questionnaire attracting more than 200 useable responses, but not large in the wider context of the entire psychology researcher population. Three important factors contributed to the limitations associated with the samples in this thesis: 1) the unavoidable aversion to statistics in many researchers, which is likely to reduce engagement with statistics-related research, 2) the sampling methods used for this research, which relied heavily on Twitter and a variety of mailing lists and therefore reached many early career researchers and members of the open science community; and 3) the ongoing COVID-19 pandemic. As the target population has been affected by ongoing university closures, rapid transitions to online teaching, and large adjustments to their own research, along with all the other burdens the pandemic has created, it is highly likely that this will have reduced the interest or time that researchers have for taking part in studies such as these.

Some of the measurement choices used throughout this thesis must also be evaluated (note that study-specific questions have been evaluated within each chapter). With regards to the demographic data, the open science variable in particular captured a very diverse group of participants, given the wording and response items: "*Are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar?"*; with options yes, no or prefer not to say. A closed question like this fails to differentiate researchers who have strong interest in the New Statistics; or more broadly have a strong familiarity with statistical reform or similar relevant activities, from those who have perhaps only once shared open data, or pre-registered their studies. In addition, statistical use was examined through the lens of self-reported behaviour, which is likely to be an inaccurate reflection of actual behaviour, particularly when compared to reviews of reporting rates in the literature.

8.6 Conclusion

This thesis has demonstrated that effect sizes, confidence intervals and statistical power are now required, or at least recommended, in order to publish articles in many of the top ranked journals in psychology. This may have a positive impact on reporting behaviour, given that participants acknowledged a lack of requirements or motivation as barriers to behaviour change. However, the evidence regarding statistical knowledge in this thesis affirms that researchers are not yet equipped with sufficient knowledge to effectively use and interpret these statistics. There is now a risk that statistical reform will be a superficial movement which represents nothing more than a series of new tick-box exercises, with little change to the integrity or reliability of psychological research. Future efforts should focus on statistical education, while also critically reflecting on which statistical changes will have a genuine impact on the reliability of research.

References

American Psychological Association. (2001). *Publication manual of the American psychological association* (5th ed.). Washington, DC.: Author.

Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). Scientists rise up against statistical significance. *Nature Comment.* https://www.nature.com/articles/d41586-019-00857-9?fbclid=IwAR1jzbGpWu9wsHIwBdOu3byOielCLEQxPZMvHJ-3X4GW2gvy4eD98a7a9EU

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist, 63*(9), 839-851. 839 https://doi.org/10.1037/0003-066x.63.9.839

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3. https://psycnet.apa.org/doi/10.1037/amp0000389

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine, 8*(1), 1-7. https://doi.org/10.1186/1741-7015-8-17

Badenes-Ribera, L., Frias-Navarro, D., Iotti, N. O., Bonilla-Campos, A., & Longobardi, C. (2018). Perceived statistical knowledge level and self-reported statistical practice among academic psychologists. *Frontiers in Psychology, 9*, 996. https://doi.org/10.3389/fpsyg.2018.00996

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from spain. *Psicothema, 27*(3), 290-295. https://doi.org/10.7334/psicothema2014.283

Badenes-Ribera, L., Frias-Navarro, D., Pascual-Soler, M., & Monterde-i-Bort, H. (2016). Knowledge level of effect size statistics, confidence intervals and meta-analysis in spanish academic psychologists. *Psicothema, 28*(4), 448-456. https://doi.org/10.7334/psicothema2016.24

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

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, Han LJ. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*(8), 1069-1077. https://doi.org/10.1177/0956797616647519

Bakker, M., Van Dijk, A., & 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

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

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389. https://doi.org/10.1037/1082-989x.10.4.389

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C., Clyde, M., Cook., De Boeck, P., Dienes, Z., Dreber, A., Easwaran., Efferson, C. … & Johnson, E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019). An introduction to second-generation p-values. *The American Statistician, 73*(sup1), 157-167. https://doi.org/10.1080/00031305.2018.1537893

British Psychological Society. (2014). Code of human research ethics.

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician, 73*(sup1), 271-280. https://doi.org/10.1080/00031305.2018.1518266

Callaway, E. (2011). Report find massive fraud at Dutch universities. *Nature,* 479, 15 (2011). https://doi.org/10.1038/479015a

Champely, S. (2017). *Pwr: Basic functions for power analysis. R package version 1.3-0.*

Cohen, J. (1990). Things I have learned (so far). . *The American Psychologist, 45*(12), 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

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

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

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. https://doi.org/10.1111/1467-8721.ep10768783

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

Collins, E., & Watt, R. (2021). Using and understanding power in psychological research: A survey study. *Collabra: Psychology, 7*(1), 28250. https://doi.org/10.1525/collabra.28250

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences, 24*(3), 200-207. https://doi.org/10.1016/j.tics.2019.12.009

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1*, 26. https://doi.org/10.3389/fpsyg.2010.00026

Counsell, A., & Harlow, L. (2017). Reporting practices and use of quantitative methods in canadian journal articles in psychology. *Canadian Psychology/psychologie Canadienne, 58*(2), 140. https://doi.org/10.1037/cap0000074

Crooks, N. M., Bartel, A. N., & Alibali, M. W. (2019). Conceptual knowledge of confidence intervals in psychology undergraduate and graduate students. *Statistics Education Research Journal, 18*(1), 46–62

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals and meta-analysis* (1st ed.). New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the american psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*(3), 138-146. https://doi.org/10.1111/j.1742-9536.2011.00037.x

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N. & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*(3), 230-232. https://doi.org/10.1111/j.1467-9280.2007.01881.x

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11*(3), 217. https://doi.org/10.1037/1082-989x.11.3.217

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*(4), 299-311. https://doi.org/10.1207/s15328031us0304_5

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*(7), 757. https://doi.org/10.1037/bul0000154

Drisko, J., & Maschi, T. (2015). *Content analysis*. Oxford University Press.

Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. R. (2006). Effect size reporting in applied psychology: How are we doing? *The Industrial-Organizational Psychologist, 43*(4), 29-37.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One, 5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. https://doi.org/10.1007/s11192-011-0494-7

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561. https://doi.org/10.1177/1745691612459059

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62*(5), 749-770. https://doi.org/10.1177/001316402236876

Fidler, F. (2006, July 2-6). *Should psychology abandon p values and teach CIs instead? Evidence-based reforms in statistics education*. [Paper presentation]. ICOTS-7, Salvador, Brazil.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*(2), 119-126. https://doi.org/10.1111/j.0963-7214.2004.01502008.x

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C.,. Anderson, J. & Goodman, O. (2004). Reform of statistical inference in psychology: The case of memory & cognition. *Behavior Research Methods, Instruments, & Computers, 36*(2), 312-324. https://doi.org/10.3758/bf03195577

Finney, D. J. (1971). *Probit analysis: A statistical treatment of the sigmoid response curve* (3rd ed.). Cambridge: Cambridge University Press.

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. *Handbook of Research Methods in Personality Psychology,* 149-169.

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23*(1), 98-122. https://doi.org/10.1177/0959354312436870

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The interpretation of scholars' interpretations of confidence intervals: Criticism, replication, and extension of Hoekstra et al.(2014). *Frontiers in Psychology, 7*, 1042. https://doi.org/10.3389/fpsyg.2016.01042

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

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

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PloS One, 12*(4), e0175583. https://doi.org/10.1371/journal.pone.0175583

Goodman, S., & Greenland, S. (2007). Assessing the unreliability of the medical literature: A response to "why most published research findings are false". Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 135. http://biostats.bepress.com/jhubiostat/paper135

Green, P., & MacLeod, C. (2019). *Simr: Power analysis for generalised linear mixed models by simulation.*

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science, 5*(1), 171511. https://doi.org/10.1098/rsos.171511

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research, 7*(1), 1-20.

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*(6), 1033-1037. https://doi.org/10.3758/BF03213921

Hoekstra, R., Johnson, A., & Kiers, H. A. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72*(6), 1039-1052. https://doi.org/10.1177/0013164412450297

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

Hoekstra, R., Morey, R. D., & Wagenmakers, E. (2018, July 8-13). *Improving the interpretation of confidence and credible intervals*. [Paper presentation]. ICOTS-10, Kyoto, Japan.

Horton, R. (2015). Offline: What is medicine's 5 sigma. *Lancet, 385*(9976), 1380. https://doi.org/10.1016/S0140-6736(15)60696-1

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681. https://doi.org/10.1177/00131640021970808

International Committee of Medical Journal Editors. (2021). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals.*

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics, 15*(1), 1-12. https://doi.org/10.1093/biostatistics/kxt007

JASP Team. (2021). *JASP (version 0.16)*

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kalinowski, P. (2010, July 11-16). *Identifying misconceptions about confidence intervals*. [Paper presentation]. ICOTS-8, Ljubljana, Slovenia.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions, 26*(3), 258-287. https://doi.org/10.1177/0163278703255242

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137. https://doi.org/10.1037/a0028086

Kerns, S., Kim, H. A., Grinspoon, E., Germine, L., & Wilmer, J. (2020). Toward a science of effect size perception: The case of introductory psychology textbooks. *Journal of Vision, 20*(11), 1185-1185. https://doi.org/10.1167/jov.20.11.1185

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350-386. https://doi.org/10.3102/00346543068003350

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education, 69*(3), 280-309. https://doi.org/10.1080/00220970109599489

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. (1st ed.) American Psychological Association.

Krell, F. (2010). Should editors influence journal impact factors? *Learned Publishing, 23*(1), 59-62. https://doi.org/10.1087/20100110

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes, 8*(1), 1-9. https://doi.org/10.1186/s13104-015-1020-4

Lakens, D. (2014). Observed power, and what to do if your editor asks for post-hoc power analyses. Posted to http://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html

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

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics,* , 363-374.

Lavine, M. (2019). Frequentist, bayes, or other? *The American Statistician, 73*(sup1), 312-318. https://doi.org/10.1080/00031305.2018.1459317

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition, 21*, 1.

Lyu, X., Xu, Y., Zhao, X., Zuo, X., & Hu, C. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology, 14* https://doi.org/10.1017/prp.2019.28

Lyu, Z., Peng, K., & Hu, C. (2018). P-value, confidence intervals and statistical inference: A new dataset of misinterpretation. *Frontiers in Psychology, 9*, 868. https://doi.org/10.3389/fpsyg.2018.00868

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

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*(sup1), 235-245. https://doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806.

Miller, J., & Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review, 23*(1), 124-130. 10.3758/s13423-015-0859-7

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49*(1), 103-120. https://doi.org/10.1111/j.1744-6570.1996.tb01793.x

Morey, R. (2018). Redefining statistical significance: The statistical arguments. Posted to https://richarddmorey.medium.com/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. (2016a). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103-123. 10.3758/s13423-015-0947-8

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E. (2016b). Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychonomic Bulletin & Review, 23*(1), 131-140. https://doi.org/10.3758/s13423-015-0955-8

Morris, P. H. (2020). Misunderstandings and omissions in textbook accounts of effect sizes. *British Journal of Psychology, 111*(2), 395-410. https://doi.org/10.1111/bjop.12401

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E., Ware, J. J. & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631. https://doi.org/10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the journal of consulting and clinical psychology. *Journal of Consulting and Clinical Psychology, 78*(3), 287. https://doi.org/10.1037/a0019294

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics, 3*(4), 201-230. https://doi.org/10.1207/s15328031us0304_1

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

Peng, C. J., Chen, L., Chiang, H., & Chiang, Y. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25*(2), 157-209. https://doi.org/10.1007/s10648-013-9218-2

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. https://doi.org/10.3758/pbr.16.2.225

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19. https://doi.org/10.1037/1082-989x.13.1.19

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Sheth, B., & Patel, J. (2015). Human perception of statistical significance and effect size. *Journal of Vision, 15*(12), 337-337. https://doi.org/10.1167/15.12.337

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR network. *BMC Medicine, 8*(1), 1-6. https://doi.org/10.1186/1741-7015-8-24

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

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*(9), 160384. https://doi.org/10.1098/rsos.160384

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

TARG Meta-Research Group. (2020). Statistics education in undergraduate psychology: A survey of UK course content. PsyArXiv. 10.31234/osf.io/jv8x3

The Jamovi Project. (2021). *Jamovi (version 1.6)*

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology, 6*, 726. https://doi.org/10.3389/fpsyg.2015.00726

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology, 67*(5), 1037-1040. https://doi.org/10.1080/17470218.2014.885986

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Smíra, M., Epskamp, S., Matzke, D., Rouder., J. & Morey, R. (2018). Bayesian inference for psychology. part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57. https://doi.org/10.3758/s13423-017-1343-3

Watt, R., and Collins, E. (2019). *Statistics for psychology* (1st ed.). London: SAGE Ltd.

Wicklin, R. (2017,). Fisher's transformation of the correlation coefficient. Message posted to https://blogs.sas.com/content/iml/2017/09/20/fishers-transformation-correlation.html

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594. https://doi.org/10.1037/0003-066X.54.8.594

Yanai, I., & Lercher, M. (2019). Night science. *Genome Biology, 20* https://doi.org/10.1186/s13059-019-1800-6

Yuan, K., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141-167. https://doi.org/10.3102/10769986030002141

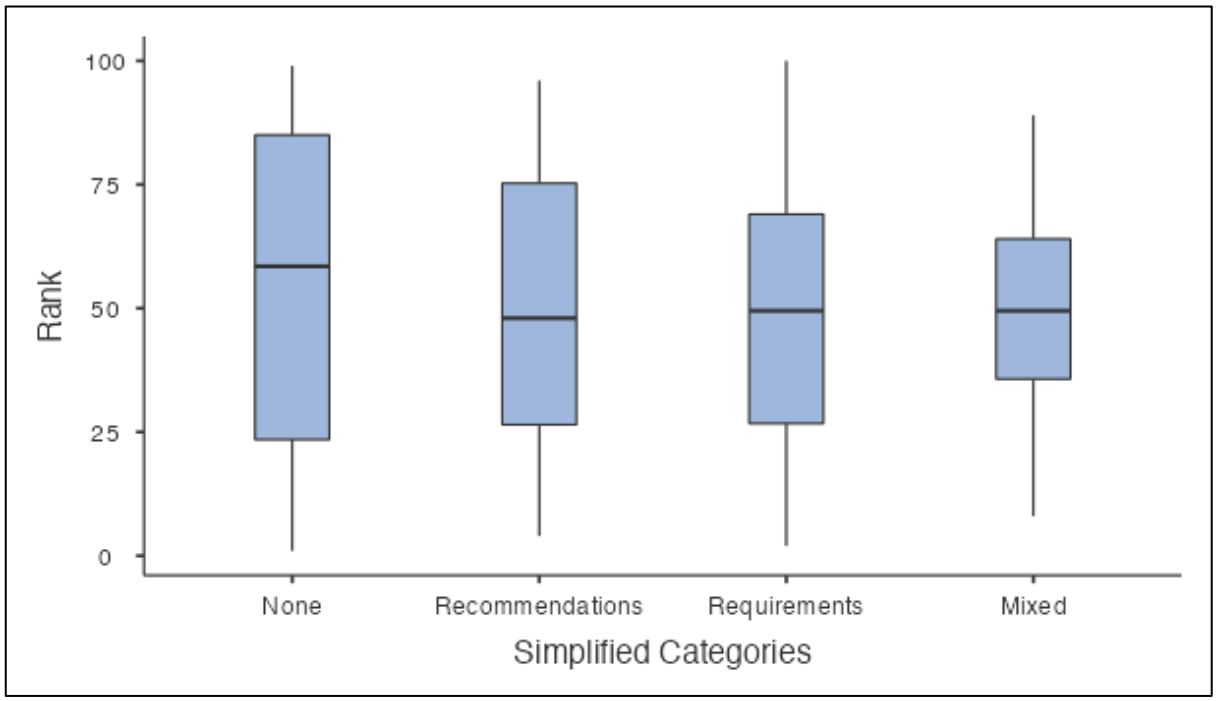
Appendix A

Appendix A contains the additional content for Chapter 2. Appendix A-1 presents the coding strategy used to review journal guidelines. Appendices A2 and A3 show that ranking position (1-100) and impact factor did not noticeably differ for each category of journal guidelines (none, recommendations, requirements, or mixed). Appendix A4 presents an overview of the full journal database associated with Chapter 2.

Appendix A1: Coding Strategy

Code	Explanation
New Statistics	
No	No mention of statistical concept (or related organisational guidelines)
Rec	Statistical concept recommended explicitly by journal
Rec – G	Journal recommends adhering to guidelines which include statistic
Req	Statistical concept required explicitly by journal
Req - G	Journal requires adherence to guidelines which include statistic
Mixed	Requirements or recommendations differ by article type
Sample Size	
No	No mention of sample size (or related organisational guidelines)
Rec	Sample size justification recommended explicitly by journal
Rec – G	Recommended to follow guidelines which mention sample size
Req	Sample size justification required explicitly by journal
Req - G	Required to follow guidelines which mention sample size
Mixed	Requirements or recommendations differ by article type
Power	
No	No mention of statistical power (or related organisational guidelines)
Req - PA	Authors are required to compute a priori power analyses
Req - S	Authors are required to compute sensitivity analyses
Req – any	Authors are required to compute either a priori or sensitivity analysis
Req – RR	Authors are required to compute power analysis for registered reports
Req – JARS	Authors are required to follow JARS (includes power)
Req - discuss	Authors are required to acknowledge statistical power
Rec – JARS	Authors are advised to follow JARS (includes power)
Rec - discuss	Journal explicitly recommends that statistical power is acknowledged
Rec – PA	Journal explicitly recommends that power analysis is used & reported
NHST	
No	No mention of *p*-values or NHST (or related organisational guidelines)
Exact	Must report exact *p*-values
Rec - G	Recommended to follow guidelines with advice about NHST
Req – G	Required to follow guidelines with advice about NHST
Other	Other rules or comments about NHST
Other	Type of guidelines mentioned, link to organisational guidelines provided (yes/no), presence or absence of statistical resources, any other notes

Appendix A2: Rankings (1-100) Split by Category



Appendix A3: Journal Impact Factor Details

Category	***Mean (SD) IF***	***Minimum IF***	***Maximum IF***
None	5.92 (3.86)	3.61	20.7
Recommendations	5.21 (1.51)	3.65	11.3
Requirements	5.66 (2.53)	3.60	13.7
Mixed	5.14 (1.59)	3.82	8.98

Appendix A4: Journal Summary Database

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
1	International Review of Sport and Exercise Psychology	20.652	None	None	None	None	None
2	Nature Human Behaviour	13.663	Req	Req	Req	None	Req
3	BEHAVIORAL AND BRAIN SCIENCES	12.579	None	None	None	None	None
4	PSYCHOLOGICAL METHODS	11.302	Rec	Rec	Rec	Rec	Rec
5	AMERICAN PSYCHOLOGIST	10.885	Req	Req	Req	Req	Req
6	LEADERSHIP QUARTERLY	10.517	Req	Req	Rec	None	None
7	European Journal of Psychology Applied to Legal Context	9.3	None	None	None	None	None
8	JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY	8.982	Mix	Mix	Mix	None	Mix
9	JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY	7.673	Req	Req	Req	Req	Req
10	JOURNAL OF APPLIED PSYCHOLOGY	7.429	Req	Rec	Req	Rec	Rec
11	Journal of Occupational Health Psychology	7.25	None	None	None	None	None
12	Clinical Psychological Science	7.169	None	Rec	Rec	None	None
13	PERSONNEL PSYCHOLOGY	7.073	Rec	Rec	Rec	Rec	Rec
14	PSYCHOLOGICAL SCIENCE	7.029	Req	Req	Req	Rec	Req
15	COMPUTERS IN HUMAN BEHAVIOR	6.829	Rec	Rec	Rec	None	None
16	JOURNAL OF BUSINESS AND PSYCHOLOGY	6.76	Rec	Rec	Rec	None	Rec

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
17	CLINICAL PSYCHOLOGY-SCIENCE AND PRACTICE	6.724	Mix	Mix	Mix	Mix	Mix
18	JOURNAL OF ABNORMAL PSYCHOLOGY	6.673	None	None	None	None	None
19	PSYCHOTHERAPY	6.596	Rec	Rec	Rec	Rec	Rec
20	Developmental Cognitive Neuroscience	6.464	Rec	Rec	Rec	None	None
21	Body Image	6.406	Rec	Rec	Rec	None	Rec
22	WORK AND STRESS	6.357	None	None	None	None	None
23	Behavior Research Methods	6.242	Rec	Rec	Rec	Rec	Rec
24	ENVIRONMENT AND BEHAVIOR	6.222	None	None	None	None	None
25	CHILD DEVELOPMENT	5.899	None	Rec	None	Rec	Rec
26	EUROPEAN JOURNAL OF PERSONALITY	5.838	Req	Req	Req	Rec	Req
27	JOURNAL OF EDUCATIONAL PSYCHOLOGY	5.805	Rec	Rec	Rec	Rec	Rec
28	COGNITIVE BEHAVIOUR THERAPY	5.761	Rec	Rec	Rec	None	None
29	AUTISM	5.689	Req	Req	Req	None	Req
30	EUROPEAN PSYCHOLOGIST	5.569	None	None	None	None	None
31	PSYCHONONEMIC BULLETIN & REVIEW	5.536	Rec	Rec	Rec	Rec	Rec
32	International Journal of Clinical and Health Psychology	5.35	Rec	Rec	Rec	None	None
33	JOURNAL OF CONSULTING AND CLINICAL PSYCHOLOGY	5.348	Req	Req	Req	Req	Req
34	Autism Research	5.216	Mix	Mix	Mix	None	Rec

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
35	JOURNAL OF ENVIRONMENTAL PSYCHOLOGY	5.192	Req	Req	Req	Rec	Req
36	JOURNAL OF THE LEARNING SCIENCES	5.171	None	None	None	None	None
37	DEVELOPMENTAL SCIENCE	5.131	None	None	None	None	None
38	PSYCHOLOGICAL ASSESSMENT	5.123	Rec	Rec	Rec	Rec	Rec
39	JOURNAL OF PERSONALITY	5.117	Req	Req	Req	None	Req
40	Psychosocial Intervention	5.083	Req	Req	Req	Req	Req
41	JOURNAL OF CLINICAL CHILD AND ADOLESCENT PSYCHOLOGY	4.964	Mix	Req	Req	Rec	Req
42	ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES	4.941	Rec	Rec	Rec	None	Rec
43	JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL	4.913	None	Rec	Rec	Rec	Rec
44	PSYCHOLOGY OF SPORT AND EXERCISE	4.785	Rec	Rec	Rec	Rec	Rec
45	BRITISH JOURNAL OF SOCIAL PSYCHOLOGY	4.691	Rec	Rec	Rec	Rec	Rec
46	JOURNAL OF COUNSELING PSYCHOLOGY	4.685	Req	Req	Req	Req	Req
47	Mindfulness	4.684	Rec	Rec	Rec	None	Rec
48	ASSESSMENT	4.667	None	None	None	Mix	Rec
49	JOURNAL OF OCCUPATIONAL AND ORGANIZATIONAL PSYCHOLOGY	4.561	None	Rec	None	None	None
50	EUROPEAN EATING DISORDERS REVIEW	4.52	Rec	Rec	Rec	None	Rec
51	BEHAVIOUR RESEARCH AND THERAPY	4.473	Mix	Mix	Mix	None	Mix
52	HUMAN DEVELOPMENT	4.452	Mix	Mix	Mix	None	Mix

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
53	Social Psychological and Personality Science	4.451	Req	Req	Req	Req	Req
54	JOURNAL OF YOUTH AND ADOLESCENCE	4.381	Rec	Rec	Rec	None	Rec
55	PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN	4.376	Req	Req	Req	Req	Req
56	Psychology of Aesthetics Creativity and the Arts	4.349	None	None	None	None	None
57	POLITICAL PSYCHOLOGY	4.333	None	None	None	None	None
58	SCHOOL PSYCHOLOGY *'QUARTERLY'*	4.333	Req	Req	Req	Req	Req
59	EMOTION	4.329	Req	Req	Req	Req	Req
60	Journal of Mental Health	4.299	None	None	None	None	None
61	CURRENT PSYCHOLOGY	4.297	Rec	Rec	Rec	None	Rec
62	JOURNAL OF SCHOOL PSYCHOLOGY	4.292	Rec	Rec	Rec	None	None
63	CONTEMPORARY EDUCATIONAL PSYCHOLOGY	4.277	Rec	Rec	Rec	None	None
64	BRITISH JOURNAL OF PSYCHOLOGY	4.267	None	Rec	None	None	None
65	HEALTH PSYCHOLOGY	4.267	Req	Req	Req	Req	Req
66	Sport Exercise and Performance Psychology	4.25	Req	Rec	Rec	Rec	Req
67	Journal of Positive Psychology	4.197	None	None	None	None	None
68	BEHAVIOR THERAPY	4.183	Mix	Mix	Mix	None	Mix
69	EVOLUTION AND HUMAN BEHAVIOR	4.178	None	None	None	None	None
70	Cyberpsychology Behavior and Social Networking	4.157	Req	Req	Req	None	Rec

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
71	Psychology of Violence	4.147	Req	Req	Req	Rec	Req
72	BRITISH JOURNAL OF CLINICAL PSYCHOLOGY	4.125	None	Rec	None	None	None
73	JOURNALS OF GERONTOLOGY SERIES B-PSYCHOLOGICAL SCIENCES AND SOCIAL SCIENCES	4.077	Rec	Rec	Rec	Rec	Req
74	PSYCHOLOGY OF WOMEN QUARTERLY	4.062	None	None	None	None	None
75	CORTEX	4.027	Rec	Rec	Rec	None	None
76	PSYCHOPHYSIOLOGY	4.016	Rec	Rec	Rec	Rec	Rec
77	European Journal of Work and Organizational Psychology	3.968	Rec	Rec	Rec	None	None
78	PSYCHOLOGY AND PSYCHOTHERAPY-THEORY RESEARCH AND PRACTICE	3.915	Rec	Rec	Rec	None	None
79	ADDICTIVE BEHAVIORS	3.913	None	None	None	None	None
80	PSYCHO-ONCOLOGY	3.894	Mix	Rec	Rec	None	Rec
81	PSICOTHEMA	3.89	Rec	Rec	Rec	None	Rec
82	INTERNATIONAL PSYCHOGERIATRICS	3.878	Req	Rec	Rec	Rec	Req
83	Journal of Happiness Studies	3.852	None	None	None	None	None
84	DEVELOPMENTAL PSYCHOLOGY	3.845	Rec	Rec	Rec	Rec	Req
85	JOURNAL OF ABNORMAL CHILD PSYCHOLOGY	3.837	Rec	Rec	Rec	None	Rec
86	International Journal of Mental Health and Addiction	3.836	Rec	Rec	Rec	None	Rec
87	ATTACHMENT & HUMAN DEVELOPMENT	3.833	Rec	Rec	Rec	None	None
88	Journal of Applied Research in Memory and Cognition	3.83	Req	Req	Rec	None	Req

Rank	Journal Name	2020 Impact Factor	NHST	ES	CI	Power	Sample Size
89	MEDIA PSYCHOLOGY	3.824	None	Rec	None	Mix	None
90	LAW AND HUMAN BEHAVIOR	3.795	Req	Req	Req	Mix	None
91	JOURNAL OF PERSONALITY ASSESSMENT	3.777	None	None	None	None	None
92	PSYCHOTHERAPY RESEARCH	3.768	None	None	None	None	None
93	APPLIED PSYCHOLOGY-AN INTERNATIONAL REVIEW-PSYCHOLOGIE APPLIQUEE-REVUE INTERNATIONALE	3.712	None	None	None	None	None
94	JOURNAL OF APPLIED BEHAVIOR ANALYSIS	3.695	None	None	None	None	None
95	JOURNAL OF GAMBLING STUDIES	3.655	Rec	Rec	Rec	None	Rec
96	COGNITION	3.65	Rec	Rec	Rec	None	None
97	INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES	3.632	None	None	None	None	None
98	Personality Disorders-Theory Research and Treatment	3.623	None	None	None	Req	Req
99	Journal of Managerial Psychology	3.614	None	None	None	None	None
100	JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY	3.603	Req	Req	None	Req	Req

Appendix B

Appendix B contains the additional content for Chapter 3 (Effect Size Survey), which includes ethical approval evidence, the study adverts and questionnaire, and a correlation matrix to evaluate true-false scale validity.

Appendix B1: Ethical Approval Evidence for Chapter 3

UNIVERSITY of STIRLING

Elizabeth Collins
Faculty of Natural Sciences
University of Stirling
FK9 4LA

General University Ethics Panel (GUEP)
University of Stirling
Stirling
FK9 4LA
Scotland UK

E: GUEP@stir.ac.uk

28 January 2020

Dear Elizabeth

Re: Examining Effect Size Knowledge, Attitudes and Experiences– GUEP 829

Thank you for submitting the minor amendments for the above proposal to the General University Ethics Panel. The ethical approaches of this project have now been approved by Chairs Action.

Please ensure that your research complies with University of Stirling policy on storage of research data which is available at:
https://www.stir.ac.uk/about/professional-services/information-services-and-library/current-students-and-staff/researchers/research-data/plan-and-design/our-policy/

If you have not already done so, I would also strongly encourage you to complete the Research Integrity training which is available at: https://canvas.stir.ac.uk/enroll/CJ43KW

If you have any concerns or queries, please do not hesitate to contact the Committee by email to guep@stir.ac.uk.

Yours sincerely,
Pp
Claire Exley
On behalf of GUEP
Dr William Munro
Deputy Chair of GUEP

Appendix B2: Example Advertising Tweet

"Are you doing/teaching quant research in psych (inc. PhD students)? Please take part in my new PhD study about effect sizes. Suitable for all – even if you don't really know what effect sizes are (we really want to hear from you!). Worldwide participants welcome. Link: *XXX*."

Appendix B3: Email Advert Template

Dear all,

Are you a psychology researcher (in any form, including PhD students and teaching-focused staff)? If you use quantitative statistics in your research, or teach quantitative statistics, then this email is for you! Please consider taking the time to participate in this survey about effect sizes. If you are unsure about what effect sizes are, we want to hear from you. If you use them all the time, we want to hear from you. If you're somewhere in between, we want to hear from you too!

This study may result in free online training materials being made available to help psychologists in their own research.

Link here: *XXX*

Please feel free to share this email with other psychology researchers who use quantitative methods.

Many thanks for your time,

Elizabeth Collins

Appendix B4: Questionnaire Used for Chapter 3

[page 1]

Please use the box below to write a definition of the term 'effect size'. You are welcome to make a guess if you don't really know, or write "I don't know" in the box. [free-text response box]

[page 2]

Your next task is to answer 'true' or 'false' to each of the following six statements. There may or may not be an equal number of true/false responses. If you don't know an answer, use the 'I don't know' option instead of guessing.

Statements
Effect sizes express the magnitude of the influence one variable has on another
A larger sample size will result in a larger (stronger) effect size
If you are doing high quality research, your aim is to find the largest effect sizes possible
A statistically significant *p*-value will result in a medium or large effect size
A small effect size indicates that the null hypothesis should fail to be rejected

[page 3]

We would like to know a little more about you and your experiences with effect sizes. You are welcome to leave any of the following questions blank.

1. Do you currently calculate effect sizes in the data analysis stage of your research? [yes/no/not always]

> *If 'yes' is selected:* Which software do you use to do so? If you use more than one, please mention them all. [free-text response box]
>
> *If 'no' or 'not always' are selected:* Why do you not, or not always, choose to calculate effect sizes? [free-text response box]

2. Have you included effect sizes in any pre-prints or published papers? [yes/no/prefer not to say]
If 'yes' is selected: Did you include effect sizes due to personal preference, or due to journal/funding/institution requirements? [personal preference/requirements]

3. How important do you feel effect sizes are in psychological research? [very important/somewhat important/not very important/not important at all/I don't know]

3a. Please use the box below if you are happy to briefly explain your response. [free-text response box]

4. Do you feel that you have been provided with sufficient training in effect sizes? [yes/no/prefer not to say]

5. Would you make use of training on effect sizes if it were to be made available? [yes/maybe/no]

If 'No'/'Maybe' is selected: You responded 'no'/'maybe' to this question. Would you like to explain why? You are welcome to leave this blank. [free-text response box]

[page 4]

You have reached the end of today's survey. Before you go, we would appreciate a very small amount of demographic information from you. You may leave any or all of these questions blank.

1. Which of the following titles best applies to your current position? Please note that this is a list using the most common job positions in the UK academic market, where "Professor" is equivalent to being a tenured, top-level academic employee in other countries.

[ ] Research or Teaching employee without a PhD
[ ] PhD student or equivalent
[ ] Post-doctoral researcher
[ ] Lecturer or Senior Lecturer (or equivalent)
[ ] Professor (or equivalent level role e.g. tenured employee)
[ ] Other (please indicate a different response below, if you wish)
[ ] Prefer not to say

2. Are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, SIPS, or anything similar?

[ ] yes
[ ] no
[ ] prefer not to say

3. Which sub-field of psychology do you identify with? E.g. sports psychology, health psychology, neuropsychology. [free-text response box]

[*free text response space*]

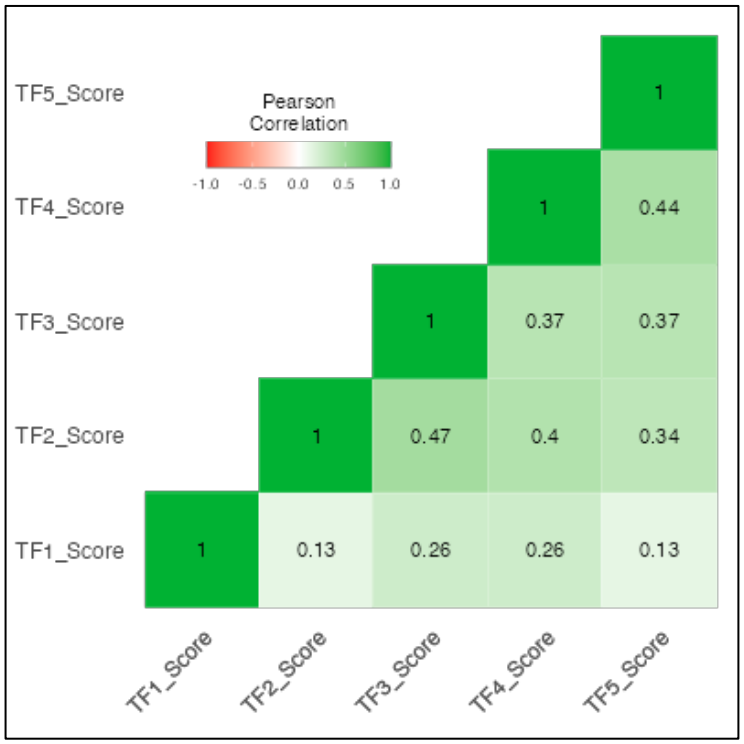
4. What country are you based in for your work or education? Please treat England, Wales, Scotland and Northern Ireland as four separate locations. [free-text response box]

Please click next for the final page of this study.

Appendix B5: Correlation Matrix for Scale Items

The 5x5 correlation matrix indicates the highest correlations are between Statement 2 and Statement 3, Statement 2 and Statement 4, and Statement 4 and Statement 5 (shown here for easy reference). Cronbach's alpha is 0.69 for the full scale, or 0.73 with statement 1 removed.

Statement	True or False?
Effect sizes express the magnitude of the influence one variable has on another	True
A larger sample size will result in a larger (stronger) effect size	False
If you are doing high quality research, your aim is to find the largest effect sizes possible	False
A statistically significant *p*-value corresponds to a medium or large effect size	False
A small effect size indicates that the null hypothesis should fail to be rejected	False



Appendix C

Appendix C contains the additional content for Chapter 4 (Effect Size Graph Study), which includes ethical approval evidence, and four graphs which show the results in Chapter 4 with Fisher's z-transformation applied. Note that this is to demonstrate that there are no critical changes in findings between the raw and transformed data.

Appendix C1: Ethical Approval Evidence for Chapter 4

UNIVERSITY of STIRLING

General University Ethics Panel (GUEP)
University of Stirling
Stirling
FK9 4LA
Scotland UK

E: GUEP@stir.ac.uk

Elizabeth Collins
Faculty of Natural Sciences
University of Stirling
FK9 4LA

18 November 2020

Dear Elizabeth,

Re: Pilot Study Evaluating Data on Graphs– GUEP (20 21) 1039

Thank you for your submission of the above to the General University Ethics Panel.

The ethical approaches of this project have been approved by GUEP, and you can now proceed with your research.

Please ensure that your research complies with University of Stirling policy on storage of research data which is available at:
https://www.stir.ac.uk/about/professional-services/information-services-and-library/current-students-and-staff/researchers/research-data/plan-and-design/our-policy/

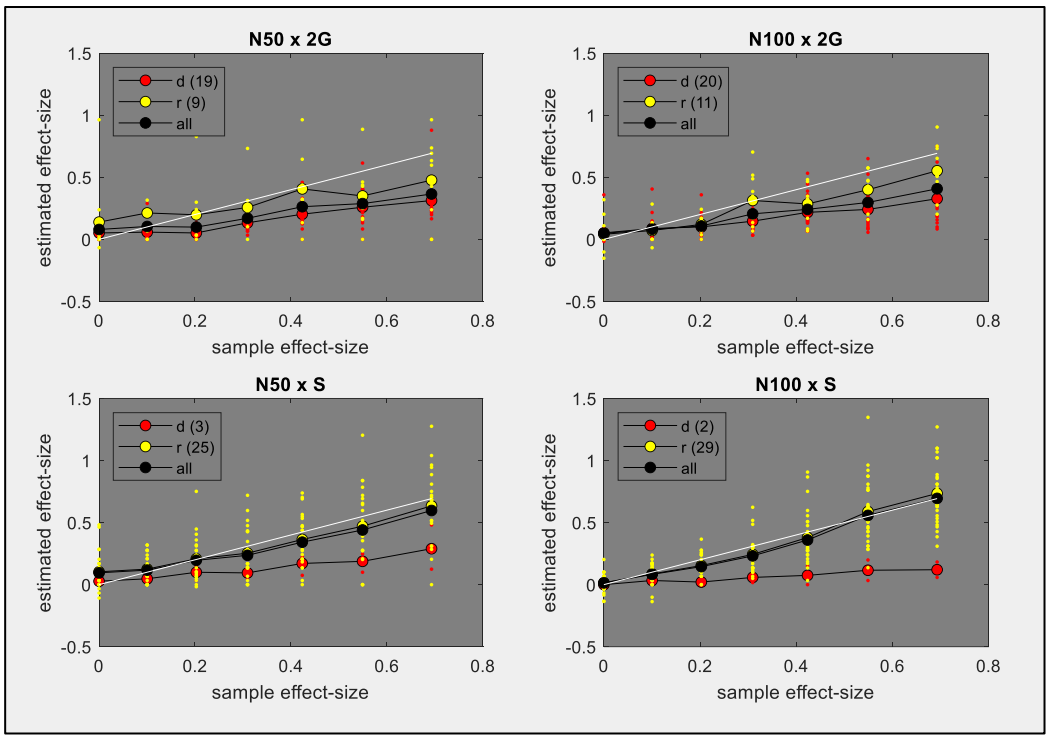
If you have not already done so, I would also strongly encourage you to complete the Research Integrity training which is available at: https://canvas.stir.ac.uk/enroll/CJ43KW

Please note that should any of your proposal change, a further submission (amendment) to GUEP will be necessary.

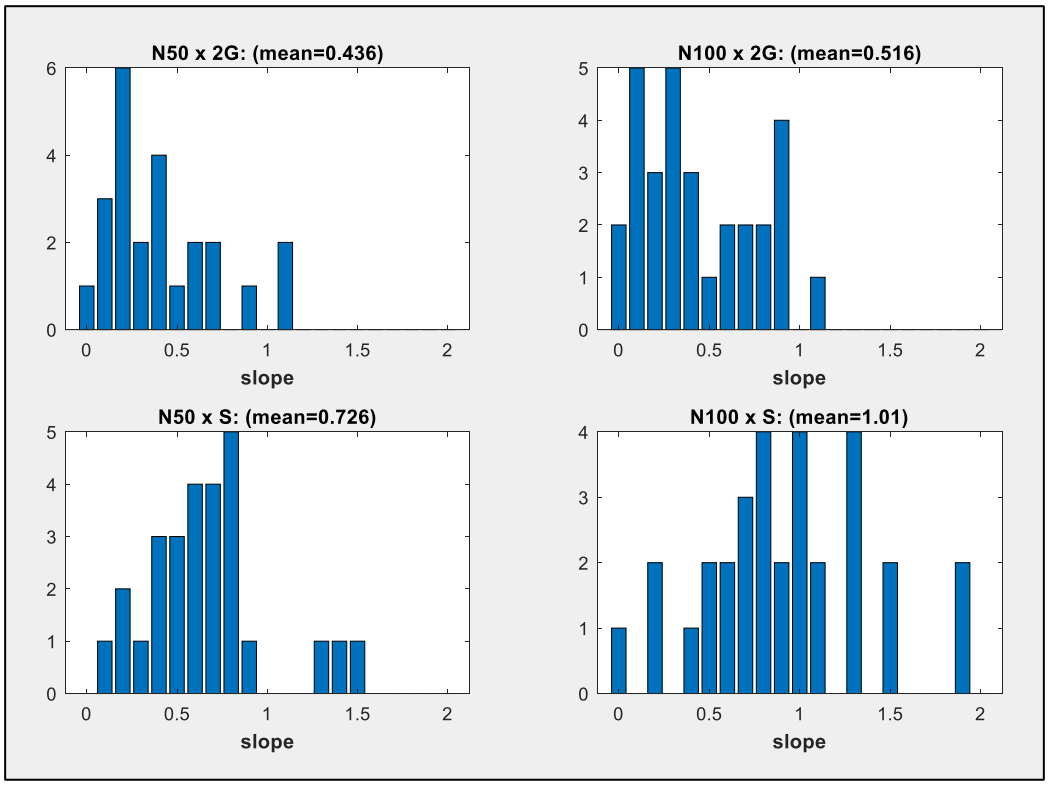
If you have any further queries, please do not hesitate to contact the Committee by email to guep@stir.ac.uk.

Yours sincerely,
Pp
Claire Exley
On behalf of GUEP

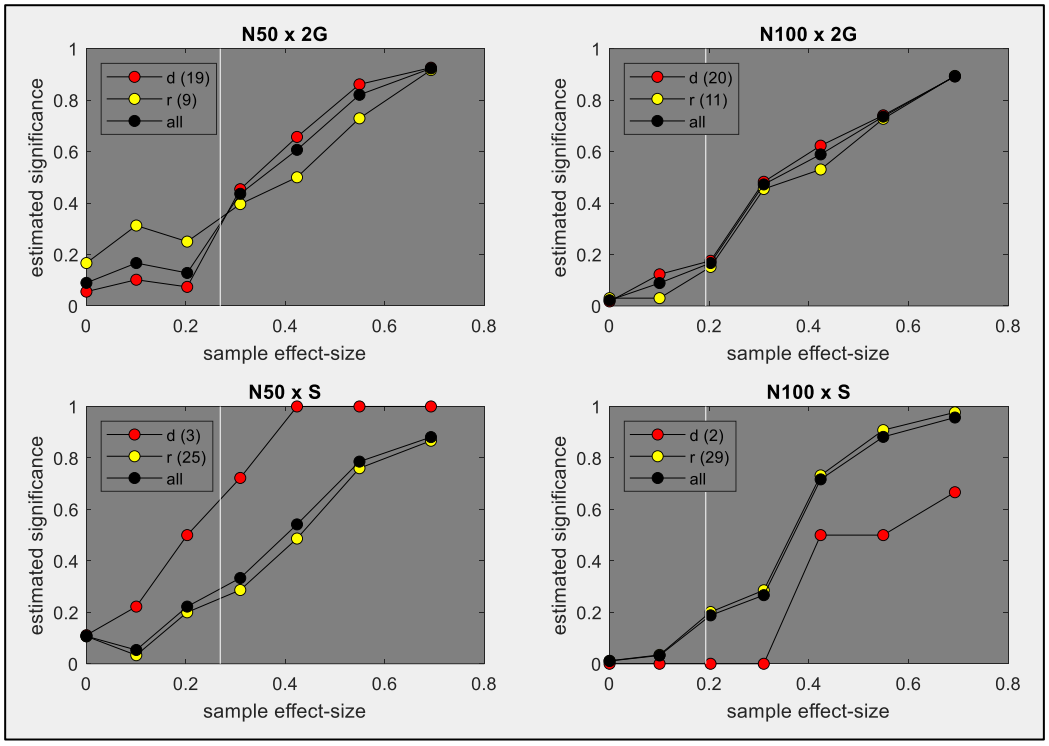
Appendix C2: Transformed Version of Figure 4.3



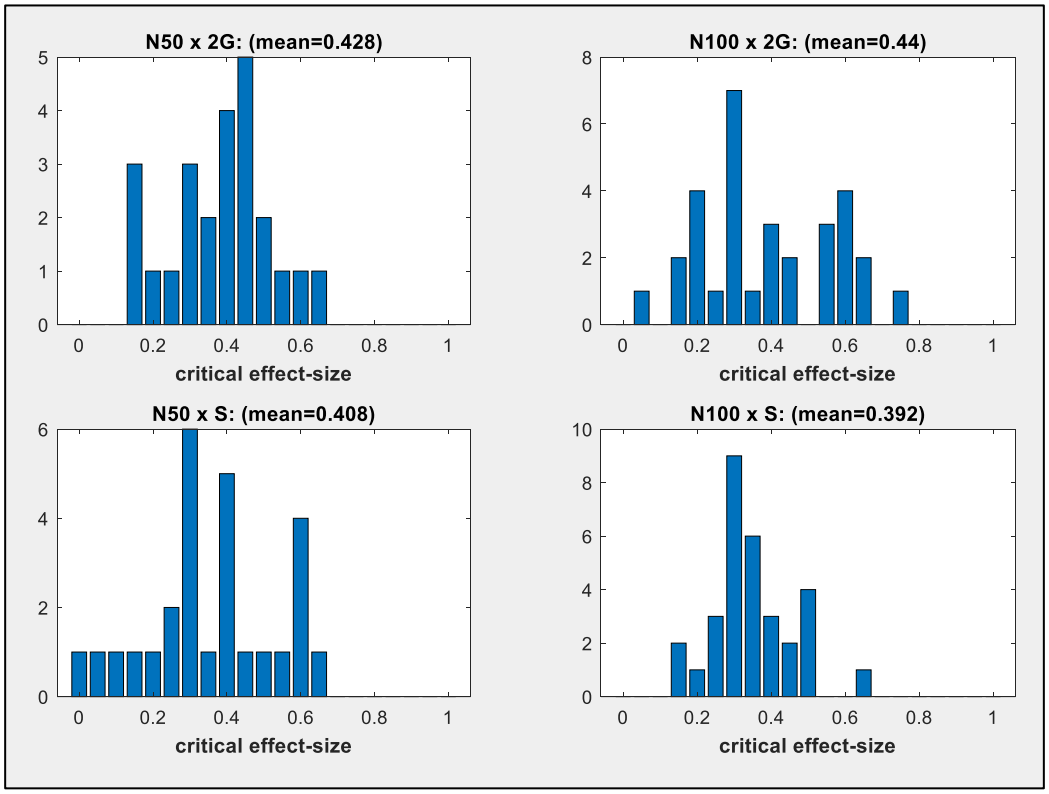
Appendix C3: Transformed Version of Figure 4.4



Appendix C4: Transformed Version of Figure 4.6



Appendix C5: Transformed Version of Figure 4.7



Appendix D

Appendix D contains the additional content for Chapter 5 and Chapter 6 (Confidence Interval Questionnaire), which includes ethical approval evidence, the questionnaire, and a correlation heatmap for evaluating the validity of the true-false scale.

Appendix D1: Ethical Approval Evidence for Chapter 5 and 6

UNIVERSITY of STIRLING

General University Ethics Panel (GUEP)
University of Stirling
Stirling
FK9 4LA
Scotland UK

E: GUEP@stir.ac.uk

Elizabeth Collins
Faculty of Natural Sciences
University of Stirling
FK9 4LA

01 April 2020

Dear Elizabeth

Re: Examining Confidence Interval Knowledge and Experiences– GUEP (19 20) 857

Thank you for your submission of the above to the General University Ethics Panel.

The ethical approaches of this project have been approved by GUEP and, providing you can carry out your research without face-to-face interactions with participants you can now proceed with your research. If your research can be moved to online or telephone interaction please follow the advice available at: https://www.stir.ac.uk/research/research-ethics-and-integrity/covid-19-guidance-for-researchers/

If your research requires face-to-face interaction then you must wait until you are advised by the University that the research recommence before you begin any data collection in person. When the project starts please inform GUEP of the amended start and end dates, there will be no requirement to seek additional approval.

Please ensure that your research complies with University of Stirling policy on storage of research data which is available at:
https://www.stir.ac.uk/about/professional-services/information-services-and-library/current-students-and-staff/researchers/research-data/plan-and-design/our-policy/

If you have not already done so, I would also strongly encourage you to complete the Research Integrity training which is available at: https://canvas.stir.ac.uk/enroll/CJ43KW

Please note that should any of your proposal change other than changes relating to the move to online participation, a further submission (amendment) to GUEP will be necessary. If you have any further queries, please do not hesitate to contact the Panel by email to guep@stir.ac.uk.

Yours sincerely,
Pp
Claire Exley
On behalf of GUEP
Professor Iain MacRury
Deputy Chair of GUEP

Appendix D2: Questionnaire Used for Chapter 5 and Chapter 6

[page 1]

1. In your own words, please define the term '95% confidence interval'. You are welcome to put 'I don't know' in this box or leave it blank and move on to answering questions about your experiences. [free-text response box]

[page 2]

We would like to know a little more about you and your experiences with confidence intervals. You are welcome to leave any of the following questions blank.

1. Do you currently calculate confidence intervals when analysing data? [yes/no/not always]

> *If 'yes' is selected:* Which software do you use to do so? If you use more than one, please mention them all. [free-text response box]
>
> *If 'no' or 'not always' are selected:* Why do you not, or not always, choose to calculate confidence intervals? [free-text response box]

2. Have you included confidence intervals in any pre-prints or published papers?

> *If 'yes' is selected:* Did you include effect sizes due to personal preference, or due to journal/funding/institution requirements? If you use confidence intervals regularly, think about which option was responsible for your first use of them. [personal preference/requirements]

3. How important do you feel confidence intervals are in psychological research? [very important/somewhat important/not very important/not important at all/I don't know]

> 3a. Please use the box below if you are happy to explain your response. [free-text response box]

[*free text response space*]

4. Do you feel that you have been provided with, or had access to, sufficient training in confidence intervals? [yes/no/prefer not to say]

5. Would you make use of training on confidence intervals if it were to be made available? [yes/maybe, but only if the training is accessible/no]

> *If 'No' is selected:* You responded 'no' to this question. Would you like to explain why? You are welcome to leave this blank.

[page 3]

Please read the following scenario and use the free text box to summarise the results (using as many or as few words as you wish). How could the confidence interval be interpreted? You are welcome to say "I don't know" or leave this question blank.

> **A study (n=42) reports that the mean weight loss and 95% confidence interval for a longitudinal diet plan is 4.65kg (-1.95, 11.25).** [free-text response box]

[page 4]

Please read the following scenario and use the free text box to summarise the results (imagine you are writing one or more sentences in a paper or discussing the results with a colleague). You are welcome to say "I don't know" or leave this question blank.

> **Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Both Skinner (2018) and Miller (2019) used two independent equal-sized groups and reported the difference between means for the group assigned to the new treatment and the group who maintained their existing treatment.**
>
> **Skinner (2018) with total n=44 found the new treatment found the difference in means was 3.61 (95% CI: 0.61 to 6.61). The study by Miller (2019) with total n=36 found that the difference in means was 2.23 (95% CI: -1.41 to 5.87). A positive difference indicates a positive outcome for the new treatment.**
>
> [free-text response box]

[page 5]

Your final task is to answer 'true' or 'false' to each of the following statements, all of which are in the context of **confidence intervals for a mean**. There may or may not be an equal number of true/false responses. If you don't know an answer, please use "I don't know".

Statements:

	True	**False**	**IDK**
A 95% confidence interval is the range of values for which you are 95% confident that the population mean falls within.			
If all other factors are held constant, an 80% confidence interval will be wider than a 95% confidence interval.			
If all other factors are held constant, a confidence interval from a sample of n=25 will be wider than a confidence interval from a sample of n=100.			
A confidence interval gives you the range of plausible values for the true sample mean.			
If an experiment is replicated with new samples from the same population, 95% of future means will fall within the original 95% confidence interval.			
If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean			

If you have any comments about your responses, or questions about the statements, please express them in the box below. You are welcome to leave this box blank.

[*free text response space*]
[page 6]

6. Thank you for sharing some information about your confidence interval experiences!

Please use the box below if you have any questions about confidence intervals. Please note that these questions won't receive a response, but are instead used to identify the most frequent questions that

researchers have, which will allow for better training resources to be developed. An overview of these questions will be disseminated by the researcher on Twitter (@Lizzie_Psych) in the future. [free-text response box]

Please use the box below if you have any other comments or thoughts about confidence intervals that you would like to share (e.g. barriers to using or understanding them, personal experiences, opinions, and so on). [free-text response box]

[page 7]

You have reached the end of today's survey. Before you go, we would appreciate a very small amount of demographic information from you. You may leave any or all of these questions blank.

1. Which of the following titles best applies to your current position?

[ ] Research or Teaching employee without a PhD
[ ] PhD student or equivalent
[ ] Postdoctoral researcher
[ ] Lecturer or Senior Lecturer (or equivalent)
[ ] Professor (or equivalent senior level role e.g. tenured employee)
[ ] Other (please indicate a different response below, if you wish)
[ ] Prefer not to say

2. Are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar?

[ ] yes
[ ] no
[ ] prefer not to say

3. Which sub-field of psychology do you identify with? E.g. sports psychology, health psychology, neuropsychology. [free-text response box]

4. What country are you based in for your work or education? Please treat England, Wales, Scotland and Northern Ireland as four separate locations. [free-text response box]

Please click next for the final page of this study.

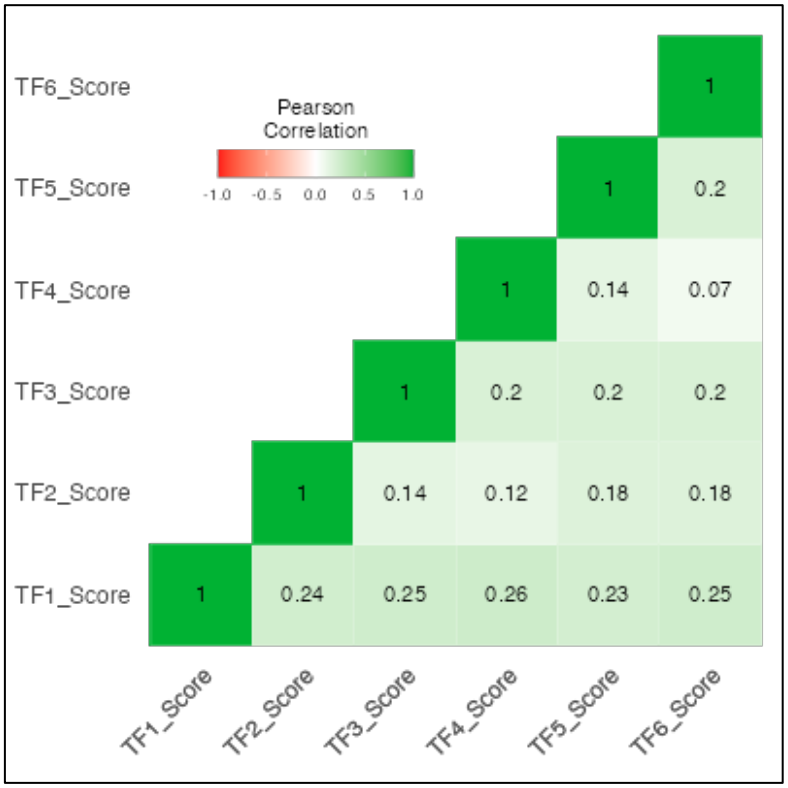
[study end page with thank you message and reminder of contact details and right to withdraw data]

Appendix D3: Correlation Matrix for Chapter 5 True-False Scale

Note that Cronbach's alpha for this scale is 0.586, and further reliability analyses indicated that removing any item would decrease Cronbach's alpha. The statements are listed here in a table for easy reference.

1	A 95% confidence interval is the range of values for which you are 95% confident that the population mean falls within.
2	If all other factors are held constant, an 80% confidence interval will be wider than a 95% confidence interval.

3	If all other factors are held constant, a confidence interval from a sample of n=25 will be wider than a confidence interval from a sample of n=100.
4	A confidence interval gives you the range of plausible values for the true sample mean.
5	If an experiment is replicated with new samples from the same population, 95% of future means will fall within the original 95% confidence interval.
6	If you repeatedly take a sample of size n from a population and construct a 95% confidence interval each time, 95% of those intervals will contain the population mean



Appendix E

Appendix E contains the additional content for Chapter 7 (Power Study), which includes ethical approval evidence and the questionnaire distributed to participants.

Appendix E1: Ethical Approval Evidence for Chapter 7

UNIVERSITY of STIRLING

General University Ethics Panel (GUEP)
University of Stirling
Stirling
FK9 4LA
Scotland UK

E: GUEP@stir.ac.uk

Elizabeth Collins
Faculty of Natural Sciences
University of Stirling
FK9 4LA

30 March 2020

Dear Elizabeth

Re: Using Power Analysis in Psychological Research– GUEP (19 20) 864

Thank you for your submission of the above to the General University Ethics Panel.

The ethical approaches of this project have been approved by GUEP and, providing you can carry out your research without face-to-face interactions with participants you can now proceed with your research. If your research can be moved to online or telephone interaction please follow the advice available at: https://www.stir.ac.uk/research/research-ethics-and-integrity/covid-19-guidance-for-researchers/

If your research requires face-to-face interaction then you must wait until you are advised by the University that the research recommence before you begin any data collection in person. When the project starts please inform GUEP of the amended start and end dates, there will be no requirement to seek additional approval.

Please ensure that your research complies with University of Stirling policy on storage of research data which is available at:
https://www.stir.ac.uk/about/professional-services/information-services-and-library/current-students-and-staff/researchers/research-data/plan-and-design/our-policy/

If you have not already done so, I would also strongly encourage you to complete the Research Integrity training which is available at: https://canvas.stir.ac.uk/enroll/CJ43KW

Please note that should any of your proposal change other than changes relating to the move to online participation, a further submission (amendment) to GUEP will be necessary. If you have any further queries, please do not hesitate to contact the Panel by email to guep@stir.ac.uk.

Yours sincerely,
Pp
Claire Exley
On behalf of GUEP

Appendix E2: Questionnaire Used for Chapter 7

Q1. As a researcher, how do you generally determine your sample size?

[free-text response box]

If power analysis is not detected as part of the response, participants will then be asked "have you ever used power analysis?". Their answer (yes/no) will set the survey flow to show questions from block A or block B.

Question Block A (participants who mention power analysis, or who answer 'yes' to previously using power analysis):

1. How frequently do you use power analysis, as a percentage of the studies you conduct that use hypothesis testing? [slider response from 0-100]

If the selection is not 100%, the following question will be displayed:

1a. Why do you not use power analysis 100% of the time? You are welcome to leave this box blank if you prefer. [free-text response box]

2. What software do you use for power analysis? You are welcome to provide multiple responses. [free-text response box]

3. How would you define "power"? You are welcome to say 'I don't know'. [free-text response box]

4. How do you establish a predicted effect size for any power analysis that you run? You are welcome to select multiple responses.

1	Use an effect size from the results of other published literature
2	Use the same effect size as a previous similar study reported in their methods
3	Use a small or medium effect size e.g. Cohen's recommendations
4	Take recommendations from other researchers
5	Use the smallest effect size of interest for my field or "meaningful" effect size for my field
6	Run a pilot study to calculate an effect size first
7	Other

5. What is the *minimum* power for studies in psychological research that you consider to be acceptable? [slider from 0 to 100%]

6. What is the *minimum* power for studies in psychological research that you would consider to be ideal? [slider from 0 to 100%]

6. How do you think opportunity sampling would affect power? Please explain any ideas you have. You are welcome to say "I don't know" or skip this question. [free-text response box]

7. How do you think outliers would affect power? Please explain any ideas you have. You are welcome to say 'I don't know' or skip this question. [free-text response box]

8. Have you ever calculated post hoc power? [yes/no/prefer not to say]

If 'yes' is selected: Why did you calculate post hoc power? If you have carried it out for more than one purpose, please mention all reasons. You are welcome to leave this box blank if you prefer. [free-text response box]

Question Block B (participants who do not mention power analysis and report 'no' to whether they have used power analysis:

1. Why have you not used power analysis in your research? All explanations are welcome, but you may leave this box blank if you prefer. [free-text response box]

2. Have you ever calculated post hoc power? [yes/no/prefer not to say]

> *If 'yes' is selected:* Why did you calculate post hoc power? If you have carried it out for more than one purpose, please mention all reasons. You are welcome to leave this box blank if you prefer. [free-text response box]

3. How would you define "power"? You can write "I don't know" in this box, or indicate that you are guessing at a definition, if you are unsure. [free-text response box]

Question Block C (all survey participants)

1. How important do you feel power analysis is in psychological research?

[ ] Not important at all
[ ] Not very important
[ ] Somewhat important
[ ] Very important
[ ] I don't know

2. On a scale from 1 (poor) to 10 (excellent), how would you describe your statistical knowledge? [slider from 1 to 10]

Demographics:

1. Which of the following titles best applies to your current position?

[ ] Research or Teaching employee without a PhD
[ ] PhD student or equivalent
[ ] Post doc
[ ] Lecturer or Senior Lecturer (or equivalent)
[ ] Professor (or equivalent level role e.g. tenured employee)
[ ] Other (please indicate a different response below, if you wish)
[ ] Prefer not to say

2. Are you actively engaged with any elements of the current movement towards improving psychological science, such as replication, pre-registration, new statistics, producing open data, the Society for the Improvement of Psychological Science (SIPS), or anything similar? [yes/no/prefer not to say]

3. Which sub-field of psychology do you identify with? E.g. sports psychology, health psychology, neuropsychology. [free-text response box]

4. What country are you based in for your work or education? If you are based in the United Kingdom, please treat England, Wales, Scotland and Northern Ireland as four separate locations. [free-text response box]

Please click next for the final page of this study.